ADAPTATIONS TO A MICROBIAL WORLD: EVOLUTIONARY OUTCOMES OF HOST-MICROBE INTERACTIONS IN *DROSOPHILA MELANOGASTER*

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Angela Michelle Early

August 2014

ADAPTATIONS TO A MICROBIAL WORLD: EVOLUTIONARY OUTCOMES OF HOST-MICROBE INTERACTIONS IN *DROSOPHILA MELANOGASTER*

Angela Michelle Early, Ph. D.
Cornell University 2014

Organisms are in constant contact with both harmful and benign microbes. Evolutionary approaches can enrich our understanding of these interactions and provide insight into their dynamics through time and across space. Here, I present an evolutionary study of the fruit fly *Drosophila melanogaster* to investigate the multiple ways microbes and parasites have shaped the evolution of this model host.

Chapter 1 explores *D. melanogaster*'s interactions with its gut bacteria. Using 37 inbred fly lines, I found that fly genotypes differ in their amount of gut bacteria. Gut microbiome size correlated with other phenotypes assayed in these lines, suggesting that commensal bacterial load may influence aspects of fly fitness—from nutrient allocation to mating behavior. While the fly only transiently interacts with these gut microbes, it maintains a lifelong relationship with the endosymbiont *Wolbachia pipientis*. In Chapter 2, I present a phylogenetic analysis of 65 globally distributed *Wolbachia* and mitochondrial genomes. *Wolbachia* infections showed strong geographic structuring and no evidence of horizontal transmission or recombination. Demonstrating a tight evolutionary relationship between host and bacteria, I

determined that all extant *Wolbachia* infections in *D. melanogaster* are monophyletic, coalescing to a single infected individual approximately 2200 years ago.

Chapter 3 more broadly considers all classes of parasites, pathogens, and commensals. Leveraging our extensive knowledge of *D. melanogaster* gene function, I infer global variation in pathogen-induced selection pressures, and find that immune processes differ in extent and route of local adaptation. Parasitoid wasps and viruses have most profoundly impacted the recent evolution of *D. melanogaster* immune genes, but the underlying genetic architectures of these adaptive events differ. Genes also experience intra-cellular selection pressures. In Chapter 4, I investigate how these intra-organismal forces shape immune gene adaptation by calculating metrics of network position and pleiotropy for each *D. melanogaster* immune gene. I found that protein-protein interactions constrain a gene's adaptive potential, but that this constraint is most apparent in processes that experience strong directional selection.

Taken together, these studies provide a more complete picture of the multi-faceted nature of host-microbe interactions, and establish an expanded framework for future research in *Drosophila* immunity.

BIOGRAPHICAL SKETCH

Angela Early arrived at Cornell University in the fall of 2007 as a graduate student in the Field of Ecology & Evolutionary Biology. Her PhD research was conducted in the laboratory of Dr. Andrew Clark and was partially supported by a National Science Foundation Graduate Research Fellowship and a Cornell University Presidential Life Sciences Fellowship. Prior to Cornell, Angela attended Yale University, where she earned a BS in Molecular, Cellular, and Developmental Biology in 2003 and an MM in Violin Performance in 2005. She then spent two years as a Science Assistant in the Division of Environmental Biology at the National Science Foundation where she first developed her interests in population genetics and host-pathogen interactions.

To my husband Damien Mahiet, whose total support is exemplified by the fact that he may be the only person to ever read this page (twice).

ACKNOWLEDGMENTS

It would be impossible to name everyone who has enriched my time at Cornell and helped me to grow as a scientist. I therefore offer here the names of those who most directly impacted the work I present in this dissertation.

My advisor Andy Clark encouraged me to think broadly, inspired me to think deeply, and fostered in me the confidence to act independently.

In their own unique ways, my committee members Brian Lazzaro and Rick Harrison not only improved the contents of this dissertation but also broadened my perspective on research and science.

A friend, colleague, and mentor, Margarida Cardoso Moreira had the patience to teach me the basics of bioinformatics. In addition, our countless coffee breaks and her limitless pep talks were invaluable to me.

Since my initial summer of failed experiments, my friend and baymate Nancy Chen kept me sane, well fed, properly punctuated, and happy.

An ever-stimulating presence, Tim Connallon pushed me to think in new directions and greatly strengthened my understanding of pop gen fundamentals.

Jen Grenier not only kept the Clark lab running smoothly but also made me a much more thorough and effective bench scientist.

The EEB department—and in particular the 2007 cohort—was a fun, supportive, and inspirational intellectual family.

Finally, I would like to acknowledge Cornell's rich, multi-disciplinary environment, as exemplified by programs and organizations like 3CGP, CVG, EvoGroup, and the Presidential Life Sciences Fellowships. My years here have been an enriching and fulfilling experience.

TABLE OF CONTENTS

INTRODUCTION

The multi-faceted nature of host-microbe interactions

Until recent decades, research on host-microbe interactions centered on disease-causing agents. Science now recognizes that the lines between friend and foe, benefit and harm are much less clear. Certain bacteria can be benign in one context but harmful in another. An immune response can become pathological if exerted too strongly. A gene that protects against one pathogen can be less efficient against a second. These trade-offs affect host fitness in real time, and also influence the evolutionary trajectory of an organism. As a result, we seldom see an immune related trait with a single evolutionary optimum. Instead, we observe polymorphisms within populations and heterogeneity in genotypes' responses to different classes of pathogens.

In this dissertation, I take two approaches to explore how the fruit fly has evolved while faced with different—and sometimes opposing—selective pressures on the immune system. The first involves experiments designed to measure the extent of phenotypic variation in natural populations. Through these experiments I have tested the immune system's efficacy in resisting bacterial infection and have also measured the community of commensal bacteria in the fly gut. The second approach uses genetic polymorphism to make inferences about past selective pressures and adaptive events. Combined, these studies explore a range of interacting partners—from gut bacteria to parasitoid wasps—that have shaped the evolutionary trajectory of the fly.

The *Drosophila* innate immune response

For several decades, *Drosophila melanogaster* has served as a key model organism for the study of innate immunity. As a result, the pathways and genes involved in pathogenic defense are among the most well characterized in the fly (Lemaitre and Hoffmann 2007).

When bacteria enter the hemolymph, they trigger the systemic immune response, which includes both humoral and cellular components. In the humoral response, pattern-recognition receptors bind common bacterial structures and initiate a signaling cascade through two NF-κB pathways (Toll and IMD). This upregulates the transcription of antimicrobial peptides (AMPs), which circulate in the hemolymph, killing bacteria and fungi. The cellular response responds to bacterial, fungal, and parasitic attack through phagocytosis, encapsulation, and melanization responses. While phagocytosis is a more general response, encapsulation is a process that targets a specific class of *Drosophila* pathogen: the parasitoid wasp. *Drosophila* also maintains a robust anti-viral defense that involves RNA interference and incorporates genes from the JAK-STAT pathway.

In addition to these systemic responses, the fly has localized immune activity in the epithelial tissues of the gut, trachea, and genitalia. These localized responses are not as well characterized, but several studies have uncovered key differences between the systemic and gut-specific immune responses. Both systems rely on the production of AMPs, but in the gut, transcription of AMPs is independent of the systemic immune response (Ryu, et al. 2004; Tzou, et al. 2000). Also in the gut, the Toll pathway plays no discernible role, and the IMD pathway appears to serve only as a back-up system, producing AMPs when bacteria prove resistant to its first line of defense, Reactive Oxygen Species (ROS) (Ryu, et al. 2008).

Drosophila-associated commensal bacteria

The fly is host to a wide array of viruses, bacteria, fungi, and parasitoid wasps. Because of the fly's role as a model organism, most research centers on those that are disease-causing agents. In the last decade, increased attention has focused on the non-pathogenic bacteria that associate with *D. melanogaster*, but there are still many holes in our understanding. The first two chapters of this dissertation partially address this gap by investigating the evolutionary relationship between the fly and two different classes of benign bacteria: gut microbes and endosymbionts.

Surveys have shown that the *Drosophila* gut microbial community is relatively simple, consisting of at most 25 different phylotypes. The majority of bacteria fall into four families: Enterobacteriaceae, Acetobacteraceae, Lactobacillaceae, and Enterococcaceae (Chandler, et al. 2011; Cox and Gilmore 2007; Wong, et al. 2011). These bacteria are tolerant of low pH, making them viable in the acidic gut environment. The microbiota are largely horizontally transmitted, however, since fly larvae often eat their egg casing, a form of vertical transmission can occur through ingestion of bacteria deposited on the outside of the egg casing during oviposition (Bakula 1969). Flies show a moderate amount of differentiation in their associated gut bacterial communities, influenced by the environment and food source (Chandler, et al. 2011; Corby-Harris, et al. 2007; Cox and Gilmore 2007; Ryu, et al. 2008). Despite this observed variation, the clear conclusion that emerges from these surveys is that the bacterial assemblages are neither random nor purely environmentally determined. Because of this non-randomness, it is likely that the fly exerts some form of control over its resident microbes, but these interactions are just starting to be explored.

To date, research on *Drosophila* gut microbes has focused on community composition. In Chapter 1, I suggest that composition is just one important aspect to consider. Using a panel of wild-type inbred fly lines, I manipulated and standardized their gut microbe composition. I was then able to quantify the amount of bacterial growth in the gut relative to the size of the fly. The heritability for this trait is high and in addition, it correlates with other aspects of host fitness. These novel observations open new questions for future microbiota research in *Drosophila* and other organisms.

While *Drosophila* gut microbes are generalists, there is one bacterium that evidences a co-evolutionary relationship with the fly: *Wolbachia pipientis*. It is a vertically transmitted alpha-proteobacterium that is passed from mother to offspring through the cytoplasm of the egg. *Wolbachia* infects a range of arthropods, in which it manifests a diversity of traits, from metabolic mutualism to reproductive parasitism. In *D. melanogaster*, the phenotypic effects exerted by *Wolbachia* are mixed and include mild mating incompatibilities (Friberg, et al. 2011) but also protection against viral infection (Teixeira, et al. 2008).

To more completely explore the evolutionary relationship between *D. melanogaster* and this major endosymbiont, I conducted a phylogenetic study of 65 inbred fly lines that carried natural *Wolbachia* infections. In Chapter 2, I present the results from this study and answer several fundamental questions about the *Drosophila*-*Wolbachia* relationship. I found that *Wolbachia* phylogeny completely mirrored the fly's mitochondrial phylogeny. This implies that mitochondria and *Wolbachia* are always co-transmitted from mother to offspring, with horizontal transmission being an extremely rare occurrence. In addition, I find the first evidence that *Drosophila* genotype exerts a regulatory effect over the level of *Wolbachia* found in the fly. Together, these results highlight the tight co-evolutionary relationship between host

and symbiont and give the first phenotypic evidence that variation in *Wolbachia* titer is partially determined by the host.

Local adaptation in the *Drosophila* immune system

Faced with an array of both commensals and pathogens, it is not surprising that the immune system shows signs of strong selection—both purifying and positive (Daub, et al. 2013; Fumagalli, et al. 2011; McTaggart, et al. 2012; Quintana-Murci and Clark 2013; Sackton, et al. 2007; Waterhouse, et al. 2007). Among *Drosophila* species in the *melanogaster* group, immune genes show an elevated rate of adaptive evolution (Sackton, et al. 2010). The strength of selection, however, is not uniform across gene functions and is mainly driven by heightened selection on recognition proteins and signal modulation proteins. The strength of selection also varies among pathways: a larger than expected number of genes in the IMD pathway show signs of accelerated evolution in both *D. melanogaster* and *D. simulans* (Jiggins and Kim 2006; Obbard, et al. 2009; Sackton, et al. 2010; Schlenke and Begun 2003).

In addition to being complex, these selective forces are likely to be environment-dependent. Pathogen pressures vary among habitats, as do scores of other variables that impact physiological characteristics and in turn immune function. The variation in selection pressure leads to local adaptation and genetic divergence of populations, which leaves signatures in the genome.

In Chapter 3, I present the most extensive study of global *Drosophila* immune gene variation performed to date. I performed analyses on single genes, identifying candidates that may have led to greater local adaptation. In addition, I leveraged our extensive knowledge about the genetic architecture of the immune response to perform a pathway-based analysis.

With this approach I identified a single class of genes (encapsulation genes) that were disproportionately differentiated between populations. Because encapsulation genes are exclusively used in parasitoid wasp defense, this observation allows us to conclude that parasitoids exert large selection pressures on their *Drosophila* hosts

Such environmental selection pressures only partially drive the mode and tempo of evolution, however. Genes are also subject to intra-organismal pressures that can promote or constrain evolution. In Chapter 4, I explore two such gene-level traits that are known to shape evolutionary rates on long time scales: network structure and pleiotropy. The *Drosophila* immune system is highly pleiotropic and many genes are known to function in other biological processes. Such pleiotropic interactions have been shown to constrain evolution over long time scales (Fraser 2005; Fraser, et al. 2002; Hahn and Kern 2005; Larracuente, et al. 2008), but no work has investigated their effect on local adaptation in *Drosophila*. I find that a proxy for pleiotropy—the number of interacting partners a protein has—negatively correlates with population differentiation. This is an important extension of studies done on longer time-scales and points to the importance of considering multiple levels of selection in studies of adaptation.

Future research will be required to elucidate the genetic underpinnings of these newly described phenotypes. Taken together, the projects completed for this dissertation have already provided a more complete picture of the multi-faceted nature of host-microbe interactions, and established an expanded framework for future research endeavors.

References

Bakula M 1969. The persistence of a microbial flora during postembryogenesis of Drosophila melanogaster. Journal of Invertebrate Pathology 14: 365-374.

Chandler JA, Lang JM, Bhatnagar S, Eisen JA, Kopp A 2011. Bacterial communities of diverse Drosophila species: ecological context of a host-microbe model system. Plos Genetics 7: e1002272. doi: 10.1371/journal.pgen.1002272

Corby-Harris V, Pontaroli AC, Shimkets LJ, Bennetzen JL, Habel KE, Promislow DE 2007. Geographical distribution and diversity of bacteria associated with natural populations of Drosophila melanogaster. Appl Environ Microbiol 73: 3470-3479. doi: 10.1128/AEM.02120-06

Cox CR, Gilmore MS 2007. Native microbial colonization of Drosophila melanogaster and its use as a model of Enterococcus faecalis pathogenesis. Infect Immun 75: 1565-1576. doi: 10.1128/IAI.01496-06

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L 2013. Evidence for polygenic adaptation to pathogens in the human genome. Molecular Biology and Evolution 30: 1544-1558. doi: 10.1093/molbev/mst080

Fraser HB 2005. Modularity and evolutionary constraint on proteins. Nature Genetics 37: 351-352. doi: 10.1038/ng1530

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW 2002. Evolutionary rate in the protein interaction network. Science 296: 750-752. doi: 10.1126/science.1068696

Friberg U, Miller PM, Stewart AD, Rice WR 2011. Mechanisms Promoting the Long-Term Persistence of a Wolbachia Infection in a Laboratory-Adapted Population of

Drosophila melanogaster. Plos One 6. doi: 10.1371/journal.pone.0016448

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. Plos Genetics 7: e1002355. doi: 10.1371/journal.pgen.1002355

Hahn MW, Kern AD 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution 22: 803-806. doi: 10.1093/molbev/msi072

Jiggins FM, Kim KW 2006. Contrasting evolutionary patterns in Drosophila immune receptors. Journal of Molecular Evolution 63: 769-780. doi: 10.1007/s00239-006-0005-2

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG 2008. Evolution of protein-coding genes in Drosophila. Trends Genet 24: 114-123. doi: 10.1016/j.tig.2007.12.001

Lemaitre B, Hoffmann J 2007. The host defense of Drosophila melanogaster. Annu Rev Immunol 25: 697-743. doi: 10.1146/annurev.immunol.25.022106.141615

McTaggart SJ, Obbard DJ, Conlon C, Little TJ 2012. Immune genes undergo more adaptive evolution than non-immune system genes in Daphnia pulex. BMC Evol Biol 12: 63. doi: 10.1186/1471-2148-12-63

Obbard DJ, Welch JJ, Kim KW, Jiggins FM 2009. Quantifying adaptive evolution in the Drosophila immune system. Plos Genetics 5: e1000698. doi: 10.1371/journal.pgen.1000698

Quintana-Murci L, Clark AG 2013. Population genetic tools for dissecting innate immunity in

humans. Nat Rev Immunol 13: 280-293. doi: 10.1038/nri3421

Ryu JH, Kim SH, Lee HY, Bai JY, Nam YD, Bae JW, Lee DG, Shin SC, Ha EM, Lee WJ 2008. Innate immune homeostasis by the homeobox gene caudal and commensal-gut mutualism in Drosophila. Science 319: 777-782. doi: 10.1126/science.1149357

Ryu JH, Nam KB, Oh CT, Nam HJ, Kim SH, Yoon JH, Seong JK, Yoo MA, Jang IH, Brey PT, Lee WJ 2004. The homeobox gene Caudal regulates constitutive local expression of antimicrobial peptide genes in Drosophila epithelia. Mol Cell Biol 24: 172-185.

Sackton TB, Lazzaro BP, Clark AG 2010. Genotype and gene expression associations with immune function in Drosophila. Plos Genetics 6: e1000797. doi: 10.1371/journal.pgen.1000797

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG 2007. Dynamic evolution of the innate immune system in Drosophila. Nature Genetics 39: 1461-1468. doi: 10.1038/ng.2007.60

Schlenke TA, Begun DJ 2003. Natural selection drives Drosophila immune system evolution. Genetics 164: 1471-1480.

Teixeira L, Ferreira A, Ashburner M 2008. The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. Plos Biology 6: 2753-2763. doi: 10.1371/Journal.Pbio.1000002

Tzou P, Ohresser S, Ferrandon D, Capovilla M, Reichhart JM, Lemaitre B, Hoffmann JA, Imler JL 2000. Tissue-specific inducible expression of antimicrobial peptide genes in Drosophila surface epithelia. Immunity 13: 737-748.

Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, Dong Y, Jiang H, Kanost MR, Koutsos

AC, Levashina EA, Li J, Ligoxygakis P, Maccallum RM, Mayhew GF, Mendes A, Michel K, Osta MA, Paskewitz S, Shin SW, Vlachou D, Wang L, Wei W, Zheng L, Zou Z, Severson DW, Raikhel AS, Kafatos FC, Dimopoulos G, Zdobnov EM, Christophides GK 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316: 1738-1743. doi: 10.1126/science.1139862

Wong CN, Ng P, Douglas AE 2011. Low-diversity bacterial community in the gut of the fruitfly Drosophila melanogaster. Environ Microbiol 13: 1889-1900. doi: 10.1111/j.1462-2920.2011.02511.x

CHAPTER 1

A new dimension of host-microbe interactions: the effect of *Drosophila* genotype on gut bacterial levels

Abstract

In some organisms, it has been shown that host genotype can influence the composition of the commensal bacterial community, but this level of host control has not been detected in the model organism *Drosophila melanogaster*. Composition, however, is only one parameter describing a microbial community. Here, we test whether a second parameter—absolute level of bacteria—is a heritable trait by quantifying the growth of four commensal bacterial strains within 37 inbred lines of *D. melanogaster*. We find that *D. melanogaster* genotype exerts a large effect on microbial level within the fly. The amounts of different bacterial strains are strongly correlated, suggesting that the role of specific interactions between bacterial species and host genotype are minimal. Additionally, we find correlations between gut commensal bacteria levels and two other phenotypes measured in these flies: mating latency and glucose content. These correlations suggest that natural variation in the amount of gut bacteria may have direct fitness consequences.

Introduction

Advances in microbiome research have demonstrated the need to consider the phenotypic effects of not only environmental conditions and organismal genotype, but also microbiome composition and by extension, the complex interactions among all three players. This holobiont concept has become an established paradigm in biology and has impacted fields from physiology to evolution (Zilber-Rosenberg and Rosenberg 2008). Studies have been con-

ducted in a wide range of organisms and have uncovered relationships between commensal bacteria and a plethora of host traits from metabolism to behavior.

In the past decade, researchers have published nearly a dozen sequence-based surveys of *Drosophila*-associated microbes (reviewed in Broderick and Lemaitre 2012). These studies have taken diverse approaches and examined the effects of food source, developmental stage, and various laboratory and natural environments. Contrary to initial expectations, however, these efforts uncovered no evidence of a well-defined core microbiome at the species level (Staubach, et al. 2013; Wong, et al. 2013). Instead, the composition of the fly microbiome is strongly affected by environmental factors such as food substrate (Chandler, et al. 2011; Staubach, et al. 2013), and its maintenance is likely dependent on constant replenishment through the ingestion of environmental microbes (Blum, et al. 2013).

Despite this large environmental effect, however, only a small subset of the microbes encountered by the fly survive within the gut (Chandler, et al. 2011). This shows that *Drosophila* exerts a certain degree of selective regulation—directly or indirectly—over its microbiome composition and raises the possibility that the lack of a core species-level microbiome may reflect a degree of functional redundancy among the microbial members. Indeed, there are certain bacterial taxa that are repeatedly sampled across *Drosophila* species and habitats. These include the genera *Acetobacter*, *Lactobacillus*, *Gluconobacter,* and *Enterococcus,* which are all acid-tolerant bacteria that can survive in the gut's low pH (Chandler, et al. 2011; Corby-Harris, et al. 2007; Cox and Gilmore 2007; Staubach, et al. 2013; Wong, et al. 2011).

While *Drosophila* has no obligate gut microbes, it is apparent that flies have evolved to maximize their fitness within a microbial context. When gut microbes are removed, the resulting axenic flies are viable, but they experience fitness costs such as metabolic dysregula-

tion (Shin, et al. 2011), altered lifespan (Brummel, et al. 2004), and enhanced susceptibility to oral pathogens (Blum, et al. 2013). In addition, specific bacterial strains have been associated with a variety of processes including insulin signaling (Shin, et al. 2011), growth and development (Lee and Brey 2013; Storelli, et al. 2011), and even mating preference (Sharon, et al. 2010). The apparently loose relationship between *Drosophila* and specific microbes therefore raises an intriguing question that is relevant to a broad array of taxa (Sachs, et al. 2011). In the absence of strong co-evolutionary relationships, how do hosts optimize the benefits they derive—or at the least, minimize the harm they receive—from transient microbial partners?

One important answer to this question is likely host regulation of bacterial growth. Indeed, unchecked commensal bacterial growth is detrimental to the fly, showing that flies have optimal fitness with some intermediate level of microbiota. Aspects of the fly's gut physiology and immune response are known to play roles in this microbial regulation. First, a low pH and the presence of digestive enzymes create an environment that is inhospitable to many bacteria. Second, the peritrophic matrix, a chitinous lining in the midgut, serves as a physical barrier, blocking microbial access to the epithelium (Kuraishi, et al. 2011). Third, a gut-specific immune response places a check on bacterial proliferation through the release of reactive oxygen species (ROS; Ha, et al. 2009) and antimicrobial peptides (AMPs; Ryu, et al. 2006). While we are forming a more comprehensive picture of how these processes respond to pathogenic infection, we know less about how the gut regulates commensal bacterial communities and maintains homeostasis (Lee and Brey 2013).

We propose that one key—and hitherto uninvestigated—aspect of the fly-microbiome relationship is the relative size of the microbial community. Here we test whether fly genotype influences not the composition, but the size of the resident microbial population. We find that

this trait does vary among flies in a heritable fashion and is largely robust to different bacterial genotypes. We find that the amount of bacteria in the gut correlates with two other fitness-related traits: mating latency and glucose content. These findings suggest that both microbiota composition and absolute microbe levels play roles in shaping host phenotype.

Methods

Fly lines and bacterial stocks

We chose 37 lines from the Drosophila Genetic Reference Panel (DGRP; Mackay, et al. 2012), a set of inbred *D. melanogaster* lines sampled in Raleigh, NC, USA. To phenotype each fly line for the amount of commensal bacterial growth within its gut, we created gnotobiotic lines that contained just a single bacterial strain. Three of these strains (*Lactobacillus brevis*, *L. plantarum*, and *Acetobacter tropicalis*) were isolated from the guts of laboratory *Drosophila* stocks (Wong, et al. 2011). The fourth was a strain of *Enterococcus faecalis* that was isolated from the hemolymph of wild-caught flies (Lazzaro, et al. 2004). Prior to the microbiota manipulations, we treated all the fly lines with tetracycline to clear them of the intracellular symbiont *Wolbachia pipientis*. *Wolbachia* has no known effect on gut microbiota, but the removal of this endosymbiont facilitated our detection of gut bacteria with qPCR. All fly lines were given the same treatment regardless of their initial *Wolbachia* infection status. For seven generations, flies were maintained on standard glucose-yeast media to which we added 0.03% tetracycline. We then returned the flies to untreated media to which we added the carcasses of four dead untreated flies of the same genotype. This ensured that the vials were seeded with the flies' original microbiota. Following restoration of the natural microbial environment, flies were maintained for at least four additional generations before being used

in the gut colonization experiments. At the end of the treatment, we confirmed that *Wolbachia* had been cleared with a standard PCR targeting the *Wolbachia wsp* gene (Zhou, et al. 1998).

Creation of gnotobiotic lines

To measure the levels of bacterial growth within fly guts, we manipulated the bacterial content of our 37 DGRP lines. We raised a bacteria-free generation of each *Wolbachia*-cleared fly line by dechorionating eggs with bleach and transferring them to autoclaved media. After adult axenic flies emerged from these vials, we transferred them to 1-inch vials with 20 ml of food on the surface of which we had added approximately 4,000 colony forming units (CFUs) of one of the four commensal bacterial strains (*A. tropicalis, E. faecalis*, *L. brevis*, and *L. plantarum*). Flies were allowed to feed on this food for one day, thereby acquiring these single-species microbial populations in their gut. To prevent the added bacteria from growing excessively on the food media, the flies were transferred to sterile food after one day where they laid eggs. For our measurements, we collected progeny from this second set of vials. These flies were never in direct contact with the initial bacterial inoculum but instead acquired their microbiome through the bacteria deposited by their parents on the food media. Before taking measurements, we allowed the flies to age for 3-5 days in a fresh, autoclaved food vial. To account for unknown and uncontrolled environmental effects, we used a block design (Figure 1.1). Each fly line-bacteria combination was established in four separate vials using two distinct axenic parental sets. Through all treatments, flies were maintained at 25°C with 12 hour light-dark cycles on Bloomington media.

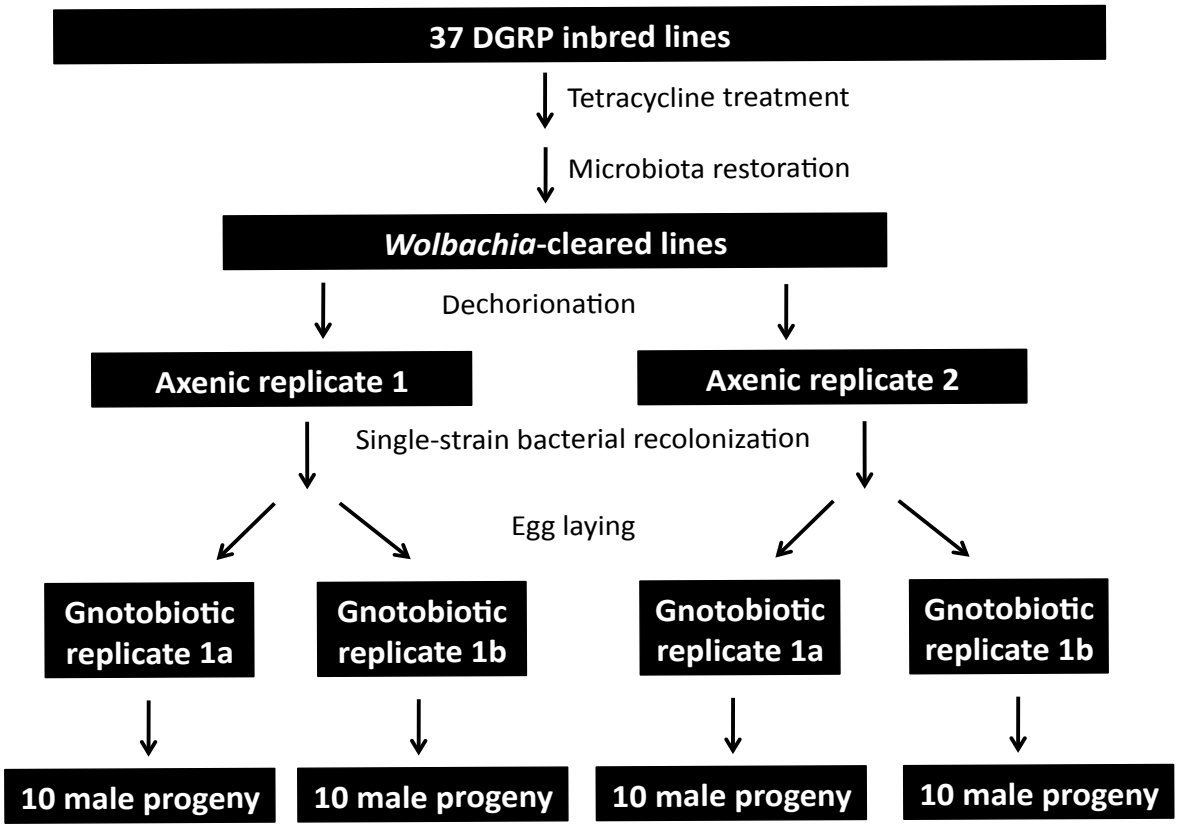


Figure 1.1. Outline of experimental design.
Both systemic pathogenic infections (left) and commensal gut recolonization (right) experiments were carried out in a block design.

Quantification of gut bacteria levels

We quantified bacterial load in the guts of the gnotobiotic fly progeny as follows. At the flies' "dawn", we transferred them to fresh, autoclaved vials. Since flies increase their feeding rate in the morning (Xu, et al. 2008), this ensured that flies ingested a minimal amount of external bacteria in the hours preceding their sampling. After 6-11 hours, flies were sexed and then washed by vortexing them for two minutes in 1.5 ml centrifuge tubes with 70% ethanol. This was followed by two 1-minute rinses in sterile water. Chandler, et al. (2011) found that three rinses adequately removed flies' surface bacteria. Flies were then immediately frozen on dry ice and maintained at -80°C until DNA extraction.

DNA was extracted from pools of 10 male flies using Qiagen DNeasy Blood and Tissue kits with a modified protocol. Briefly, flies were added to 96-well plates with 180 µl lysis buffer (20 mM Tris-Cl, 2 mM sodium EDTA, 1.2% Triton X-100, and 20 mg/ml fresh lysozyme), 200 µl Qiagen Buffer AL, four 2.0 mm zirconia beads, and 0.1 ml 0.1 mm glass beads. The plates were then processed for 2 minutes on a BioSpec Mini-Beadbeater-96. Following lysis, we added 20 µl proteinase K and incubated the samples at 56°C for 3.5 hours. To ensure there would be no remaining RNA in our sample, we performed a double RNase digest with both RNase A (10 µg/ml) and RNase T1 (25 units/ml), incubating at 37°C for 30 minutes. We then added 200 µl ethanol and proceeded with the standard Qiagen spin-column protocol.

We performed quantitative real-time PCR on total genomic DNA to determine the ratio of bacterial to fly DNA in each sample. Each 10 µl reaction contained 5 µl gDNA (approximately 30 ng) and 5 µl of Roche LightCycler 480 SYBR Green I Master. Reactions were carried out on a Roche LightCycler 480 with the following protocol: 5 minutes at 95°C followed by 50 cycles of 95°C for 15 seconds, 60°C for 30 seconds, and 72°C for 10 seconds. We measured the amount of *D. melanogaster* DNA with primers that targeted the single copy gene *Dfd* (5'-GTAGCGAAGAAACCCACCAA-3' (For), 5'-ACGCTCCACTCACCTCATTC-3' (Rev)). For each sample, we used a pair of bacteria-specific primers that provided greater sensitivity than universal bacteria primers. Primers used were: *A. tropicalis,* 5'-TAGCTAACGCGATAAGCACA-3' (For), 5'-ACAGCCTACCCATACAAGCC-3' (Rev); *E. faecalis*, 5'-TGCTTGTTGGGGTTGTAGGACTCCA-3' (For), 5'-CGGGGCTTTCACCCTCTTTAGCG-3' (Rev); *L. brevis,* 5'-TCAGTTTTGAGGGGCTTACCTCTCT-3' (For), 5'-GGCATCCACCATGCGCCCTT-3' (Rev); *L. plantarum* 5'-TGCGGCTGGATCACCTCCTTTC-3' (For), 5'-ACTGGTTCGGTTCCAATGGGCC-3' (Rev).

Under natural conditions, the fly gut would harbor a bacterial community, not a single strain, but the benefit of this simplified approach was two-fold. First, to quantify the entire community, we would have had to rely on qPCR bacterial quantification methods that target conserved regions of the 16s rRNA gene. All universal primers have variable affinities for different bacterial strains, making accurate comparisons between communities difficult. In addition, bacterial abundance is likely determined by a combination of three main factors: environment, host genotype, and microbial community composition. By maintaining a controlled environment and eliminating competition among bacterial strains, we were here able to measure the host effect in isolation.

Statistical analyses

We calculated the commensal bacterial level as the difference in the bacterial gene and *Drosophila* gene Ct values. To test the effect of fly genotype on bacterial amount, we constructed a linear mixed model using the lme4 package in R (R Development Core Team 2011). In the model, we used fly line as a fixed effect and experimental block as a random effect. To test whether there was significant variation among fly lines for their relative level of commensal bacterial, we used R to construct an ANOVA table and perform hypothesis tests on this model. We tested for correlations between commensal bacterial levels using the function cor.test in R. To calculate the broad sense heritability (H^2) for the trait, we constructed a random-effects linear model where block and line were random effects and relative commensal bacterial level was the response variable. We then calculated $H^2 = \sigma^2_G/(\sigma^2_G + \sigma^2_E)$, where σ^2_G was the variance attributed to the line effect and σ^2_E was the residual variance.

Association testing

We obtained genotype information for each of our lines from the DGRP website (DGRP Freeze 2.0; dgrp.gnets.ncsu.edu). Because we obtained phenotypes from only 37 lines, we lacked the power to perform a genome-wide association test. We therefore limited our analysis to variants found in 375 *Drosophila* genes with known immune function (Chapter 3). Using PLINK (Purcell, et al. 2007), we filtered this variant set based on minor allele frequency (MAF > 0.05) and genotyping rate (> 0.9). After filtering, we were left with 54,993 SNPs and small indels. Using this filtered set, we constructed an IBS kinship matrix with the EMMAX (Kang, et al. 2010) to control for hidden population structure. We then used EMMAX to perform association tests. As our phenotypes, we used the mean of the residuals for each line replicate as our phenotypic measure. Residuals were calculated from a linear model with line and experimental block as random effects. After identifying variants with a nominal significance of $P < 0.0001$, we annotated the functional effect of these SNPs with the Ensembl Variant Effect Predictor v. 73.

Results

Drosophila haplotypes harbor commensal bacteria populations of variable size

Using a set of 37 inbred fly lines, we tested for the presence of heritable variation in the level of commensal bacteria in the gut. For each genotype, we created gnotobiotic lines that were colonized with a single bacterial strain that is known to reside in the fly gut (*Acetobacter tropicalis*, *Enterococcus faecalis, Lactobacillus brevis*, and *L. plantarum*). Three of these strains (*A. tropicalis*, *L. brevis*, and *L. plantarum*) were directly isolated from laboratory fly

stocks and were previously shown to be dominant members of the microbial gut community in laboratory flies (Wong, et al. 2011).

Overall, we found that commensal bacteria level varies among flies in a heritable fashion (Figure 1.2). For three of our four bacteria (*A. tropicalis, L. brevis*, and *L. plantarum*), we were able to measure significant bacterial population-size differences among fly lines (Table 1.1; ANOVA, *A. tropicalis*, $P = 0.0005$, *L. brevis*, $P = 0.0008$; *L. plantarum*, $P = 0.0007$). No significant line effect was detected for our fourth bacterial strain, *E. faecalis* (ANOVA, $P = 0.462$). This bacterium was detected in only a subset of our samples (26 lines, 48 total samples), and so we likely lacked the power to make inter- and intra-line comparisons. Similarly, broad sense heritability (H^2) was substantial (>0.62) for *A. tropicalis*, *L. brevis*, and *L. plantarum*, but relatively low (0.100) for *E. faecalis* (Table 1.1).

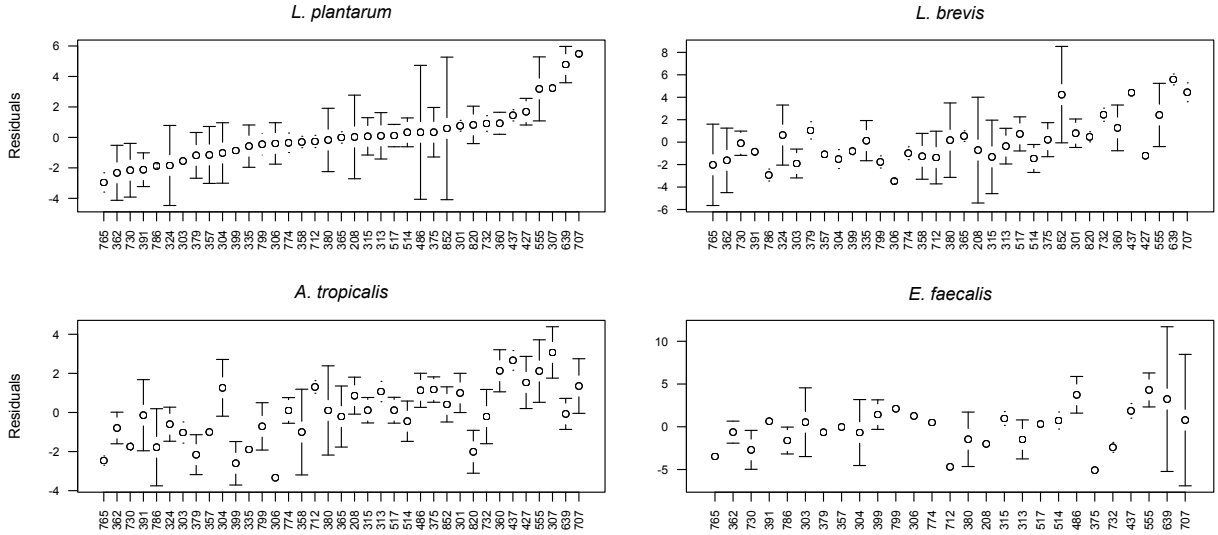


Figure 1.2. Size of commensal bacteria populations in 37 inbred fly lines.
Values shown are the residuals (± 1 S.D.) from a model accounting for block effect. Higher values correspond to a higher ratio of bacterial DNA to fly DNA as measured by quantitative PCR. In all plots, the lines are ordered according the rank order of L. plantarum residuals.

Table 1.1. Significant genetic variation and heritability in abundance of commensal gut bacteria across DGRP lines of Drosophila.

	d.f.	*F*	*P*	Genetic variance	Residual variance	Experimental variance	H^2
A. tropicalis	36	2.655	0.0005	1.382	2.223	0.05858	0.6216
E. faecalis	26	1.055	0.5	0.7888	7.882	2.157	0.1001
L. brevis	34	2.613	0.0008	2.695	4.341	0	0.6208
L. plantarum	36	2.582	0.0007	1.858	2.994	12.48	0.6206

Bacterial levels were significantly correlated in all but two comparisons, *E.faecalis-A. tropicalis* and *E. faecalis-L. brevis* (Table 1.2). The strongest correlation was between the two *Lactobacillus* species, suggesting their genetic relatedness also leads to similar responses in this assay (Pearson's $r = 0.7231$, $P = 9.29e\text{-}07$). Despite the missing data and low heritability estimates, the *E. faecalis* level still had a marginally significant correlation with *L. plantarum* (Pearson's $r = 0.4333$, $P = 0.0239$).

Table 1.2. Correlations between relative commensal bacterial levels.

	A. tropicalis	***E. faecalis***	***L. brevis***
E. faecalis	0.1126		
L. brevis	0.4590**	0.3153	
L. plantarum	0.5801***	0.4333*	0.7231***

Pearson's r. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$
Bacterial population size was calculated as the ratio of bacterial DNA to fly DNA as measured with qPCR. Correlations were performed on the residuals from a mo that accounted for batch effect.

Commensal bacterial level correlates with fitness-related phenotypes

We are beginning to understand the phenotypic effects of gut microbe composition in *Drosophila* (Newell and Douglas 2014; Sharon, et al. 2010), but the consequences of absolute

bacterial levels are yet to be explored. We were therefore interested in investigating possible connections between gut microbial levels and other fly traits. To do so, we correlated our measurements of commensal bacterial level with phenotypic measurements performed on these same lines in other studies. Previous studies on these fly lines have measured a number of phenotypes including starvation resistance (Ayroles, et al. 2009), recovery time from chill coma (Ayroles, et al. 2009), life span (Ayroles, et al. 2009), mating speed (Ayroles, et al. 2009), competitive fitness (Ayroles, et al. 2009), oxidative stress (Weber, et al. 2012), and nutritional indices (Unckless, et al. in prep). To investigate the potential fitness effects of commensal bacterial level, we looked for correlations between our data and these relevant phenotypes.

Two of these phenotypes were significantly correlated with *L. plantarum* level: mating latency and glucose content. Mating latency was measured as the time it took virgin flies of the same genotype to initiate copulation (Ayroles, et al. 2009). We found a significant, positive correlation between this trait and the level of *L. plantarum* in the gut ($\rho = 0.5664$, $P = 0.0002577$). Although there is no clear mechanism relating microbiome with mating latency, there is potential for microbiome composition to impact cuticular lipids and other olfactory cues, which can in turn affect mate preference (Sharon, et al. 2010) and mating duration (Lize, et al. 2014). Whether this trait is driven by males, females, or the interaction of the two sexes remains a question for future investigation.

In addition, we found that that level of *L. plantarum* positively correlated with levels of glucose in the fly when reared on a high glucose diet (Spearman's rank correlation test; $\rho = 0.3883$, $P = 0.03167$; Unckless et al, in prep). This correlation was even stronger with the second principal component (PC) calculated by Unckless, et al. (in prep) for a more complete set

of nutritional indices measured in these flies (ρ = -0.4988, P = 0.004756). This PC explained approximately 25% of the variance among their set of lines and had a loading of: 0.54 protein content, -0.76 glucose content, 0.06 triglyceride content, 0.17 glycerol content, and -0.3 glycogen content.

Results from association testing

The epithelial immune response is known to play a role in regulating commensal bacteria growth (Ryu, et al. 2008). It is therefore likely that segregating variation in immune-related genes might explain some of the variance we see in commensal bacterial level. Using EMMAX (Kang, et al. 2010), we performed association tests between commensal bacterial level and approximately 55,000 segregating SNPs and indels found in known immune genes. The strongest associations we found were with *L. plantarum* levels. For this phenotype there were 18 SNPs and three indels with an uncorrected p-value less than $1x10^{-4}$. In contrast, with *L. brevis* there were 3 SNPs, with *E. faecalis* there were 2 SNPs, and with *A. tropicalis* there were no SNPs that reached this significance level (Table 1.3). In total, we found significant associations with 18 genes. Four of these variants were in coding regions, four were in UTR regions, and the remaining 18 were found in introns. Interestingly, six of the genes (33%; *lozenge*, *puckered*, *hep*, *grh, Tg*, *msn*) are involved in wound repair even though wound repair genes comprise only 9% of our gene set. This suggests that aspects of epithelial integrity may play a role in mediating commensal bacteria growth.

Discussion

Drosophila has been proposed as an important model organism for studying gut physiology in general and host-microbe interactions in particular (Buchon, et al. 2013; Lee and Brey 2013).

Table 1.3. Top associations between commensal bacterial levels and segregating variants found in immune genes.

Gene	Gene ID	Chrom	Position	Effect	Bacteria	Type	*P*
Transferrin 3	FBgn0034094	2R	12117361	3' UTR	*L. plantarum*	SNP	4.19E-07
norpA	FBgn0262738	X	4224907	Intron	*L. plantarum*	SNP	7.48E-07
vir-1	FBgn0043841	2L	12421766	5' UTR	*L. plantarum*	SNP	8.84E-07
lozenge	FBgn0002576	X	9182995	Intron	*L. plantarum*	SNP	9.67E-06
puckered	FBgn0243512	3R	3936274	Intron	*L. plantarum*	SNP	9.75E-06
hep	FBgn0010303	X	12977834	Intron	*L. plantarum*	SNP	9.75E-06
hep	FBgn0010303	X	12977865	Intron	*L. plantarum*	SNP	9.75E-06
hep	FBgn0010303	X	12977966	Intron	*L. plantarum*	Indel	9.75E-06
grh	FBgn0259211	2R	13724439	Intron	*L. plantarum*	SNP	9.75E-06
grh	FBgn0259211	2R	13724846	Intron	*L. plantarum*	SNP	9.75E-06
Egfr	FBgn0003731	2R	17439269	Intron	*L. plantarum*	SNP	9.88E-06
Pvf2	FBgn0031888	2L	7072090	Intron	*L. plantarum*	SNP	1.09E-05
cher	FBgn0014141	3R	12944353	Intron	*L. plantarum*	SNP	1.74E-05
vir-1	FBgn0043841	2L	12421723	5' UTR	*L. plantarum*	Indel	1.80E-05
grh	FBgn0259211	2R	13724215	Intron	*L. plantarum*	Indel	5.66E-05
cher	FBgn0014141	3R	12927830	Intron	*L. plantarum*	SNP	6.14E-05
Tg	FBgn0031975	2L	8013307	Intron	*L. plantarum*	SNP	7.55E-05
Tg	FBgn0031975	2L	8013329	Intron	*L. plantarum*	SNP	7.55E-05
Tg	FBgn0031975	2L	8013331	Intron	*L. plantarum*	SNP	7.55E-05
Jafrac2	FBgn0040308	3L	3043681	Syn	*L. plantarum*	SNP	8.14E-05
edl	FBgn0023214	2R	14555761	Nonsyn	*L. plantarum*	SNP	9.26E-05
AGO-1	FBgn0262739	2R	9832316	3' UTR	*L. brevis*	SNP	1.39E-05
Nos	FBgn0011676	2L	10832816	Intron	*L. brevis*	SNP	8.12E-05
msn	FBgn0010909	3L	2566878	Syn	*L. brevis*	SNP	8.84E-05
RhoL	FBgn0014380	3R	5326463	Intron	*E. faecalis*	SNP	3.37E-05
mop	FBgn0036448	3L	14766523	Syn	*E. faecalis*	SNP	5.60E-05

To date, however, research efforts have focused solely on the composition or complete presence/absence of the microbial community. Here we highlight a third parameter—commensal bacterial level —and show that it is both heritable and variable in natural populations. The heritability estimates for commensal bacterial level are substantial (0.62), and for the three major bacterial strains, H^2 is higher than for several other phenotypes previously measured in these fly lines (0.25-0.58; Ayroles, et al. 2009). This demonstrates that this trait is not simply by-product of environmental conditions, but is instead a phenotype largely controlled by host genotype.

Not all variable traits impact host fitness. Here, however, we report two fitness-related phenotypes that correlate with the levels of *L. plantarum* maintained in the gut: copulation latency and glucose content. Both correlations suggest important links between microbiota levels and fly phenotypes. These correlations warrant further investigation to determine whether there is a causal link.

The positive correlation between the level of *L. plantarum* in the gut and copulation latency (as measured by Ayroles, et al. 2009) is surprising but not without precedent. Sharon, et al. (2010) found that this same bacterium influenced mate choice in *Drosophila melanogaster*. Using a single fly genotype, they determined that diet affected the relative levels of *L. plantarum* in the gut. This alteration was accompanied by a change in the composition of cuticular hydrocarbons, major players in pheromonal communication. Suggesting that bacterial levels may also affect cuticular hydrocarbons, they additionally noted that antibiotic treatment reduced the levels of cuticular hydrocarbons. These results pointed to the role of environment in shaping the downstream effects of commensal bacteria. Our new observations open the possibility that fly genotype is also at play in governing the outcomes of these host-microbe interactions.

Glucose content—the second phenotype that significantly correlated with *L. plantarum* level—is the type of metabolic trait commonly associated with gut bacteria activity. The causal mechanism behind this correlation is unknown, but given the primary role that gut bacteria play in nutrient provisioning in other organisms (Douglas 2009), it is not unlikely that the size of the commensal bacteria population influences nutrient availability and cycling in the host. The role played by absolute bacterial level is also intriguing in light of recent find-

ings that interactions between bacterial species can mediate the nutritional effect on the host (Newell and Douglas 2014).

On the host side, we found that the flies vary in the level of bacteria they contain. This effect of host genotype was relatively constant across bacteria as evidenced by strong correlations between commensal bacterial levels (Table 1.2). Recent work has shown that maintenance of the *Drosophila* gut microbiome relies on continually replenishment from environmental sources (Blum, et al. 2013). This observation, however, does not run counter to our own observations. We observe that the level of bacteria in the gut is heritable and correlates with fitness-related phenotypes. Whether the gut bacteria are reproducing in the fly or being reintroduced from the environment does not affect these observations. The need for replenishment, however, does suggest potential mechanistic bases for this trait. One is feeding rate. Flies that consume and retain a higher volume of food will likewise retain a higher level of environmental microbes within their gut. A second possibility is that commensal bacterial levels are largely determined by gut size. By comparing levels of bacterial DNA to levels of *Drosophila* DNA, our estimates of commensal bacterial level were made relative to *Drosophila* body size. While this approximates gut size, the scaling relationship between body and organ size is not static in *Drosophila* (Shingleton, et al. 2009). Third, wound repair genes were highly represented in the top hits of our SNP association test. As many of these genes have general involvement in epithelial growth and repair, these associations raise the possibility that epithelial shedding may play a role in microbe regulation. Fourth, past lab manipulations have demonstrated that the epithelial immune activity influences bacterial growth (Ryu, et al. 2008), suggesting that this may further fine-tune the degree of microbial maintenance.

The constant replenishment of gut bacteria is likely a given, as it is doubtful that flies would ever encounter a bacteria-free environment in their natural habitat. Further, when given the option of sterile food, flies prefer food containing microbiome volatiles (Venu, et al. 2014), suggesting that wild flies regulate their microbiome content through feeding and ovi-positioning preferences. Coupled with our own observations, this shows that flies have both behavioral and physiological ways of influencing their gut microbe levels and composition.

The naturally segregating variation that we observe suggests that, as with many complex traits, there is no single optimal strategy for regulating commensal bacterial levels. Previous work in *D. melanogaster* has shown that relative levels of certain bacterial strains correlate with healthy versus pathological gut states and that the immunological activity of the gut can push the fly from one condition to the other (Ryu, et al. 2008). Further investigations into the effects of the natural variation we describe here will greatly inform our understanding of host-microbiome relationships and the potential trade-offs inherent in maintaining resident microbial populations.

Acknowledgements

We wish to thank Angela Douglas and Adam Wong for helpful methodological discussions and for supplying the commensal bacterial strains used in this work. Thanks also to Jen Grenier for qPCR assistance.

References

Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RR, Mackay TF 2009. Systems genetics of complex traits in Drosophila melanogaster. Nature Genetics 41: 299-307. doi: 10.1038/ng.332

Blum JE, Fischer CN, Miles J, Handelsman J 2013. Frequent replenishment sustains the beneficial microbiome of Drosophila melanogaster. MBio 4: e00860-00813. doi: 10.1128/mBio.00860-13

Broderick NA, Lemaitre B 2012. Gut-associated microbes of Drosophila melanogaster. Gut Microbes 3: 307-321. doi: 10.4161/gmic.19896

Brummel T, Ching A, Seroude L, Simon AF, Benzer S 2004. Drosophila lifespan enhancement by exogenous bacteria. Proc Natl Acad Sci U S A 101: 12974-12979. doi: 10.1073/pnas.0405207101

Buchon N, Broderick NA, Lemaitre B 2013. Gut homeostasis in a microbial world: insights from Drosophila melanogaster. Nature Reviews Microbiology 11: 615-626. doi: 10.1038/nrmicro3074

Chandler JA, Lang JM, Bhatnagar S, Eisen JA, Kopp A 2011. Bacterial communities of diverse Drosophila species: ecological context of a host-microbe model system. Plos Genetics 7: e1002272. doi: 10.1371/journal.pgen.1002272

Corby-Harris V, Pontaroli AC, Shimkets LJ, Bennetzen JL, Habel KE, Promislow DE 2007. Geographical distribution and diversity of bacteria associated with natural populations of Drosophila melanogaster. Appl Environ Microbiol 73: 3470-3479. doi: 10.1128/AEM.02120-06

Cox CR, Gilmore MS 2007. Native microbial colonization of Drosophila melanogaster and its use as a model of Enterococcus faecalis pathogenesis. Infect Immun 75: 1565-1576. doi: 10.1128/IAI.01496-06

Douglas AE 2009. The microbial dimension in insect nutritional ecology. Functional Ecology 23: 38-47. doi: 10.1111/j.1365-2435.2008.01442.x

Ha EM, Lee KA, Seo YY, Kim SH, Lim JH, Oh BH, Kim J, Lee WJ 2009. Coordination of multiple dual oxidase-regulatory pathways in responses to commensal and infectious microbes in drosophila gut. Nat Immunol 10: 949-957. doi: 10.1038/ni.1765

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42: 348-354. doi: 10.1038/ng.548

Kuraishi T, Binggeli O, Opota O, Buchon N, Lemaitre B 2011. Genetic evidence for a protective role of the peritrophic matrix against intestinal bacterial infection in Drosophila melanogaster. Proc Natl Acad Sci U S A 108: 15966-15971. doi: 10.1073/pnas.1105994108

Lazzaro BP, Sceurman BK, Clark AG 2004. Genetic basis of natural variation in D. melanogaster antibacterial immunity. Science 303: 1873-1876. doi: 10.1126/science.1092447

Lee WJ, Brey PT 2013. How microbiomes influence metazoan development: insights from history and Drosophila modeling of gut-microbe interactions. Annu Rev Cell Dev Biol 29: 571-592. doi: 10.1146/annurev-cellbio-101512-122333

Lize A, McKay R, Lewis Z 2014. Kin recognition in Drosophila: the importance of ecology and gut microbiota. ISME J 8: 469-477. doi: 10.1038/ismej.2013.157

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RR, Barron M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu LL, Qu C, Ramia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu YQ, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D, Gibbs RA 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173-178. doi: 10.1038/nature10811

Newell PD, Douglas AE 2014. Interspecies interactions determine the impact of the gut microbiota on nutrient allocation in Drosophila melanogaster. Appl Environ Microbiol 80: 788-796. doi: 10.1128/AEM.02742-13

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575. doi: 10.1086/519795

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ryu JH, Ha EM, Oh CT, Seol JH, Brey PT, Jin I, Lee DG, Kim J, Lee D, Lee WJ 2006. An essential complementary role of NF-kappaB pathway to microbicidal oxidants in Drosophila gut immunity. EMBO J 25: 3693-3701. doi: 10.1038/sj.emboj.7601233

Ryu JH, Kim SH, Lee HY, Bai JY, Nam YD, Bae JW, Lee DG, Shin SC, Ha EM, Lee WJ 2008. Innate immune homeostasis by the homeobox gene caudal and commensal-gut

mutualism in Drosophila. Science 319: 777-782. doi: 10.1126/science.1149357

Sachs JL, Essenberg CJ, Turcotte MM 2011. New paradigms for the evolution of beneficial infections. Trends Ecol Evol 26: 202-209. doi: 10.1016/j.tree.2011.01.010

Sharon G, Segal D, Ringo JM, Hefetz A, Zilber-Rosenberg I, Rosenberg E 2010. Commensal bacteria play a role in mating preference of Drosophila melanogaster. Proc Natl Acad Sci U S A 107: 20051-20056. doi: 10.1073/pnas.1009906107

Shin SC, Kim SH, You H, Kim B, Kim AC, Lee KA, Yoon JH, Ryu JH, Lee WJ 2011. Drosophila microbiome modulates host developmental and metabolic homeostasis via insulin signaling. Science 334: 670-674. doi: 10.1126/science.1212782

Shingleton AW, Estep CM, Driscoll MV, Dworkin I 2009. Many ways to be small: different environmental regulators of size generate distinct scaling relationships in Drosophila melanogaster. Proc Biol Sci 276: 2625-2633. doi: 10.1098/rspb.2008.1796

Staubach F, Baines JF, Kunzel S, Bik EM, Petrov DA 2013. Host species and environmental effects on bacterial communities associated with Drosophila in the laboratory and in the natural environment. Plos One 8: e70749. doi: 10.1371/journal.pone.0070749

Storelli G, Defaye A, Erkosar B, Hols P, Royet J, Leulier F 2011. Lactobacillus plantarum promotes Drosophila systemic growth by modulating hormonal signals through TOR-dependent nutrient sensing. Cell Metab 14: 403-414. doi: 10.1016/j.cmet.2011.07.012

Venu I, Durisko Z, Xu J, Dukas R 2014. Social attraction mediated by fruit flies' microbiome. J Exp Biol 217: 1346-1352. doi: 10.1242/jeb.099648

Weber AL, Khan GF, Magwire MM, Tabor CL, Mackay TF, Anholt RR 2012. Genome-wide association analysis of oxidative stress resistance in Drosophila melanogaster. Plos One 7: e34745. doi: 10.1371/journal.pone.0034745

Wong AC, Chaston JM, Douglas AE 2013. The inconstant gut microbiota of Drosophila species revealed by 16S rRNA gene analysis. ISME J 7: 1922-1932. doi: 10.1038/ismej.2013.86

Wong CN, Ng P, Douglas AE 2011. Low-diversity bacterial community in the gut of the fruitfly Drosophila melanogaster. Environ Microbiol 13: 1889-1900. doi: 10.1111/j.1462-2920.2011.02511.x

Xu K, Zheng X, Sehgal A 2008. Regulation of feeding and metabolism by neuronal and peripheral clocks in Drosophila. Cell Metab 8: 289-300. doi: 10.1016/j.cmet.2008.09.006

Zhou W, Rousset F, O'Neil S 1998. Phylogeny and PCR-based classification of Wolbachia strains using wsp gene sequences. Proc Biol Sci 265: 509-515. doi: 10.1098/rspb.1998.0324

Zilber-Rosenberg I, Rosenberg E 2008. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiol Rev 32: 723-735. doi: 10.1111/j.1574-6976.2008.00123.x

CHAPTER 2

Monophyly of Wolbachia pipientis genomes within Drosophila melanogaster: Geographic structuring, titre variation, and host effects across five populations

Abstract

Wolbachia pipientis is one of the most widely studied endosymbionts today, yet we know little about its short-term adaptation and evolution. Here, using a set of 91 inbred Drosophila melanogaster lines from five populations, we explore patterns of diversity and recent evolution in the Wolbachia strain *w*Mel. Within the D. melanogaster lines, we identify six major mitochondrial clades and four *w*Mel clades. Concordant with past studies, the Wolbachia haplotypes contain an overall low level of nucleotide diversity, yet they still display geographic structuring. Using Bayesian analysis informed with demographic estimates of colonization times, we estimate that all extant D. melanogaster mitochondrial haplotypes coalesce to a Wolbachia-infected ancestor approximately 2,200 years ago. Finally, we measure *w*Mel titre within the infected flies and find that titre varies across populations, an effect attributable to host genetic factors. This demonstration of local phenotypic divergence suggests that intra-specific host genetic variation plays a key role in shaping this model symbiotic system.

Introduction

Endosymbiotic relationships are increasingly recognized as key drivers of adaptation and speciation (McFall-Ngai et al. 2013). Genomic comparisons — both among endosymbionts as well as between endosymbionts and their nearest free-living relatives — have brought to light a number of key observations about the evolution of bacterial symbionts in general and intra-

cellular symbionts in particular (Medina & Sachs 2010; Moran et al. 2008; Moya et al. 2008), but we are only beginning to understand the intraspecific variation that affects their short-term evolution (Moran et al. 2009; Richardson et al. 2012).

One of the more widely studied endosymbionts is Wolbachia pipientis, an a-Proteobacterium estimated to infect 40% of terrestrial arthropods (Zug & Hammerstein 2012) as well as some nematodes (Taylor et al. 2005). Wolbachia resides in both somatic and gonadal tissue and is transferred from mother to offspring through the egg cytoplasm. Despite this reliance on vertical transmission, however, Wolbachia evolution has been marked by frequent host-jumps (Baldo et al. 2006b; Werren et al. 1995). These large evolutionary transitions were accompanied by recombination and genomic rearrangements (Baldo et al. 2006a; Klasson et al. 2009), which may have been enabled by key genomic characteristics — in particular, the maintenance of functional DNA repair and recombinational machinery (Wu et al. 2004). Little is known, however, about the genetic factors that influence population dynamics within single Wolbachia lineages.

Wolbachia induces a range of phenotypic changes in its hosts. While acting as an obligate mutualist in some filarial nematodes, it is best known as a reproductive parasite in insects, inducing cytoplasmic incompatibility (CI), parthenogenesis, feminization, and male killing. Compared to many other Wolbachia strains, *w*Mel, the strain that infects the fruit fly Drosophila melanogaster, causes more moderate phenotypic effects. These include fitness-enhancing phenotypes such as heightened viral resistance (Teixeira et al. 2008) and increased iron tolerance (Brownlie et al. 2009), as well as low levels of CI (Friberg et al. 2011; Reynolds & Hoffmann 2002).

Developing a deeper understanding of the persistence and ecological importance of *w*Mel infections will rely on a more thorough description of the mutational processes and selection pressures that shape the bacterium's evolution. Genomic regions that are known to be variable among different Wolbachia strains have shown essentially no variation within *w*Mel. Until recently, previous analyses of global genetic diversity have been limited to a few known structural variants (Nunes et al. 2008b; Riegler et al. 2005). Importantly, these studies identified a number of divergent *w*Mel lineages and showed that the frequencies of these haplotypes dramatically changed in the latter half of the 20th century. More recently, Richardson et al. (2012) provided a first look at genome-wide *w*Mel diversity. Their study leveraged data from two different large-scale D. melanogaster sequencing efforts (the Drosophila Population Genomics Project and the Drosophila Genetic Reference Panel) that focused on multiple sparsely-sampled populations within Africa, one sparsely-sampled population within Europe, and one deeply-sampled population within North Carolina. This previous study provided key insights into *w*Mel transmission, nucleotide evolution, and depth of coverage, but the different sequencing and sampling approaches used in the two panels makes comparisons between the populations difficult. Furthermore, the African and European sequences were derived from haploid embryos, preventing phenotypic analyses of adults.

Here, we present genomic sequences of 65 *w*Mel strains from five geographically diverse populations of D. melanogaster, providing a picture of global genome-wide nucleotide diversity in this model endosymbiont. Combined with the reconstruction of mitochondrial sequences from these same fly lines, this high-resolution dataset allows us to address three aspects of recent *w*Mel evolution that are key to advancing our understanding of the D. melanogaster-*w*Mel symbiosis. First, we analyze patterns of molecular evolution in the *w*Mel genome

to provide a summary of its global genetic diversity and patterns of transmission. Second, we combine demographic information with a Bayesian phylogenetic reconstruction to estimate the date of the cytoplasmic Most Recent Common Ancestor (MRCA). Finally we examine the extent to which a key phenotype, the within-fly density of *w*Mel, is determined by genotypic differences among its D. melanogaster hosts.

Materials and Methods

Drosophila lines, DNA extraction, and sequencing

We used 91 inbred Drosophila melanogaster lines from 5 populations: Beijing (China); Ithaca, NY (USA); Netherlands; Tasmania; and Zimbabwe (the Global Diversity Lines; Table 2.1). The lines were established from isofemale lines and then inbred for 12 generations, as described in Greenberg et al. (2010). DNA was extracted from pools of 50 adult female flies using Qiagen DNeasy Blood & Tissue kits. Samples were then sequenced to approximately

Table 2.1. Geographic distribution of mtDNA and wMel haplotypes.

	Beijing (China)	Ithaca, NY (USA)	Netherlands	Tasmania	Zimbabwe	Total
I	1 (1)	19 (14)	1 (0)	4 (1)	1 (1)	26 (17)
II	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	1 (1)
III	3 (2)	0 (0)	15 (12)	15 (13)	16 (14)	49 (41)
IV	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
V	0 (0)	0 (0)	2 (0)	0 (0)	0 (0)	2 (0)
VI	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	1 (0)
VII	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
VIII	12 (6)	0 (0)	0 (0)	0 (0)	0 (0)	12 (6)
Total	16 (9)	19 (14)	19 (12)	19 (14)	18 (16)	91 (65)

For each population, the number of sampled mtDNA haplotpes within each major clade is listed, followed by the number of lines carrying a Wolbachia infection in parentheses. Haplotypes correspond with those given in Fig 2. Clades I-VI are defined by Richardson *et al.* (2012); clade VII is defined in Ilinsky (2013); clade VIII is described in this paper.

12x nuclear genomic depth at the Beijing Genomics Institute. Sequencing was performed on an Illumina HiSeq2000 using 100-bp paired-end reads with a 450-500 bp insert size (manuscript in preparation).

Read alignment and genomic variant detection

Raw reads from each D. melanogaster line were aligned to the D. melanogaster mitochondrial genome (RefSeq NC001709.1, r5.44) and to the Wolbachia pipientis strain *w*Mel genome (GenBank AE017196) using Mosaik v2.1.33 (http://bioinformatics.bc.edu/marthlab/Mosaik). Duplicate reads were marked using Picard v1.56 (http://picard.sourceforge.net). The resulting alignments were then fed through a standard Picard-GATK pipeline to call nucleotides at each site and to identify indels under five bp in length (DePristo et al. 2011; McKenna et al. 2010). Briefly, BAM files were merged and indexed with Picard, then realigned and genotyped with GATK. As the genomes were small, we used hard filtering instead of GATK's variant quality score recalibration pipeline. For the Wolbachia data set, we based our filters on GATK's best practices v.3. Because of the extremely high coverage of the mitochondrial genomes, we modified the filters for these sequences (for indels: QD < 2.0, ReadPosRankSum < -20.0, and FS > 400; for SNPs: QD<3.0, MQ<35.0, HaplotypeScore > 13.0, MQRankSum < -45, and ReadPosRankSum < -8.0). To obtain a representative Canton-S mitochondrial genome, we aligned Illumina reads from Canton-S ovaries (Sequence Read Archive, SRR353680; Soshnev et al. 2012) to the mitochondrial reference genome using filters recommended in GATK's best practices v. 3, except we set the mapping quality cutoff to 17.

A site was masked if a base call was made for fewer than 50% of the Drosophila lines or if it overlapped a GATK-called indel. Alternate allele calls were marked as missing in in-

dividual lines if the read depth in the line at that position was less than three. For the purpose of our analyses, we disregarded heterozygous calls made by GATK, calling the site based on the most frequent nucleotide at that position in that line's alignment. Essentially, this means we sampled a single strain of Wolbachia from each fly line. (Similar to how inbred fly lines sample a single chromosome from the original, wild-caught fly.) The decision to follow this procedure was three-fold. First, as discussed in the Results, we determined that no fly lines carried multiple haplotypes representative of different clades. Second, due to inbreeding, relaxed selection, and potential within-fly drift, it was unclear how heteroplasmy within inbred lab lines would inform our study of diversity in the wild. And finally, calculating diversity statistics would require the development and fitting of a complex statistical/population genetic model that would account for these "heterozygous" sites, which in the context of a Wolbachia infection arise not from diploidy but rather under a wide range of potential allele frequencies. Note that the impact of this approach was likely minimal: among the calls at variable sites in the final *w*Mel data set, GATK had made 9,419 homozygous calls and only 19 heterozygous calls. In the mtDNA dataset, these numbers were 14,965 and 19, respectively.

Pindel v.0.2.4 (Ye et al. 2009) was used to identify inversions, tandem duplications, insertions between 5 and 80 bp, and deletions of 5 bp or greater. This program uses mapping information from paired-end reads to infer the presence of structural variants. Initial filtering removed calls with only single strand support. We subsequently removed weakly supported calls that had high strand bias and low read count. Post-hoc, as a means of determining whether these filters were overly stringent, we noted that all the removed calls were incongruous with our SNP-constructed phylogenetic trees. Riegler et al. (2005) used two variable number tandem repeat loci, a large inversion, and two IS5 transposon insertion sites to distinguish

among five different Wolbachia genotypes in D. melanogaster. We compared the locations of the large inversion and IS5 insertion sites to the break points identified by Pindel.

For each genome from each line, mean read depth (mean number of reads mapped at each nucleotide position) was calculated using GATK's DepthofCoverage analysis. To create standardized estimates of Wolbachia and mtDNA density within each fly line, we calculated the ratio of aligned *w*Mel or mtDNA reads to aligned D. melanogaster nuclear genome reads (Supplementary Information, Table S1; data not shown). Nucleotide diversity (π) was calculated using custom Perl scripts.

Testing for lateral genetic transfer

To test whether portions of the *w*Mel genome have been transferred to the nuclear genome of its host, we looked for evidence of paired reads where one mate read aligned to the *w*Mel genome while the other aligned to the D. melanogaster genome. Such read pairs could suggest that a piece of the *w*Mel genome relocated to a D. melanogaster chromosome. Using SAMtools v0.1.18 (Li et al. 2009), we separated out all read pairs where one read mapped to *w*Mel while the other was unmapped. The unmapped reads were then aligned to the complete D. melanogaster reference genome (r5.46) using the Mosaik protocol outlined above.

Phylogenetic analyses

The ancestral states of *w*Mel SNPs were determined using W. pipientis strain *w*Ri (RefSeq NC_012416.1). Homologous gene regions were identified using the Ensembl database. As the mutation rate of mitochondrial DNA is high and recurrent mutation is possible, mitochondrial SNPs were polarized using 15 complete D. simulans genomes (Ballard 2000; AF200833.1-

AF200842.1, AF200844.1- AF200846.1, AF200848.1- AF200849.1). Sequences were aligned with Muscle v3.8.31 (Edgar 2004) and the ancestral state at each SNP was determined by eye. After initial tree building, we used parsimony to infer the ancestral states of *w*Mel SNPs that lacked clear *w*Ri homologs and mtDNA SNPs where the ancestral state was ambiguous due to segregating variants in D. simulans (Early & Clark 2013).

We constructed phylogenetic trees using both maximum likelihood and Bayesian methods. Maximum likelihood trees were constructed with RAxML v7.2.8 (Stamatakis 2006) using the GTRCAT method, 300 multiple inferences, and the default hill-climbing algorithm. The best-scoring maximum likelihood tree was chosen and support for the tree was calculated with 10,000 bootstrap replicates. Bayesian trees were constructed with MrBayes v3.2 (Ronquist et al. 2012) using reversible jump MCMC to estimate the number of independent substitution rate parameters (nst=mixed). We ran the MCMC analysis for 5 million generations and discarded a 25% burn-in fraction prior to analysis. Results were checked by eye in Tracer v1.5 (http://beast.bio.ed.ac.uk/Tracer) to ensure convergence. Bayesian trees were also constructed with BEAST (Drummond & Rambaut 2007) during the course of the Most Recent Common Ancestor analysis discussed below. For both methods, we included either *w*Ri or the D. simulans sequences described above in order to infer the root of the tree. In addition to these trees, we also constructed phylogenetic networks using the Neighbor-Net method in SplitsTree4 v4.12.3 (Huson & Bryant 2006). A tanglegram combining the *w*Mel and mtDNA trees was constructed with Dendroscope v3.2.3 (Huson & Scornavacca 2012). For the Neighbor-Net analysis, ambiguous sites were inferred with parsimony where possible. We removed any remaining sites with missing data prior to analysis. Unless otherwise stated, the Wolbachia trees were constructed with a concatenated sequence composed of all identified variable sites. Mi-

tochondrial analyses were performed using the first 14,916 nucleotides of the genome. After tree construction, we identified major clades and named them based on the system established in Richardson et al. (2012).

Trees with additional D. melanogaster mitochondrial haplotypes were constructed with MrBayes using the settings outlined above. In addition to the Canton-S sequence assembled from reads in SRA (see above), we downloaded the following mitochondrial sequences from GenBank: Alstonvl (FJ190106.1), Barcelona (JX266575.1), BER1 (JQ686694.1), Brownsvl (FJ190107.1), CO3 (JQ686695.1), Dahomey (FJ190108.1), Hawaii (JX266576.1), Israel (JX266577.1), Japan (FJ190109.1), Madang (JX266578.1), Mysore (FJ190110.1), Oregon R (AF200828.1), Oregon R-C (JQ686698.1), Puerto Montt (JX266579.1), QI2 (JQ686696.1), Reids1 (JQ686697.1), Sweden (JX266580.1), tko25t (JQ686693.1), w1118iso (FJ190105.1), and Zimbabwe 53 (AF200829.1). We aligned all sequences (approximately 12,300 bp) with Muscle v3.8.31 (Edgar 2004) prior to tree construction.

Node age, root age, and substitution rate estimates

To estimate the date of the most recent cytoplasmic coalescence, we calculated the divergence times of our mitochondrial haplotypes with BEAST v1.7.2 (Drummond & Rambaut 2007). To most closely approximate unconstrained, neutrally evolving sequences, we created a concatenated dataset of all third codon sites. BEAST analyses were then run with a strict molecular clock, the Hasegawa-Kishino-Yano (HKY) substitution model, no site heterogeneity, and a constant population size. The MCMC chain ran for 100 million generations and a burn-in fraction of 10% was discarded prior to analysis. We examined log files in Tracer to ensure we acquired an adequate Effective Sample Size for each parameter.

First, we estimated the age of the root by using a strict clock based on the mitochondrial mutation rate estimated by Haag-Liautard et al. (2008) (6.2×10^{-8} mutations per site per fly generation) and assuming 10 fly generations per year. Second, to incorporate demographic estimates into our analysis, we placed age priors at nodes C and D (Fig. 2; normal distribution with a mean of 200 years and a standard deviation of 50 years). The clock rate prior was set at 6.2×10^{-7} mutations per site per year with one of three standard deviations (1×10^{-7}, 1×10^{-6}, or 1×10^{-5}). All other assumptions were the same as above. For each model, we estimated the marginal maximum likelihood using both path sampling and stepping-stone analyses (Baele et al. 2012; Baele et al. 2013). Bayes Factors were calculated to choose among models.

The *w*Mel substitution rate was calculated relative to the mtDNA rate by running in parallel the same BEAST analyses with the *w*Mel sequence data partitioned into codon positions and intergenic regions. The root age and clock rates for each genomic region were then scaled by the mtDNA results.

Molecular evolutionary analyses

SNPs were functionally annotated using Ensembl's Variant Effect Predictor v2.3 (McLaren et al. 2010). Genome-wide Ka and Ks values were calculated with KaKs Calculator (Zhang et al. 2006) using the Goldman-Yang (GY) maximum likelihood method on concatenated codon-aligned coding sequences. To determine mutational bias, the total number of each nucleotide within the *w*Mel genome was counted from the *w*Mel reference genome. Similarly, mutational bias calculations were based on changes to the reference strand. Codon usage was calculated from both confirmed and predicted protein-coding sequences as annotated in the Ensembl Bacteria database, release 15 (McLaren et al. 2010). All calculations were based only on the

sites covered in our alignments. Statistical analyses were conducted in R (R Development Core Team 2011).

Quantification of w*Mel density*

We reared flies from 61 of the 65 Wolbachia-infected lines at room temperature in vials of standard glucose-yeast media. At the larval stage, we chose two replicate vials from each line, ensuring a comparable, moderate larval density across all lines. Pools of twenty mated females, aged 6-8 days, were chosen from each vial. Flies were ground and DNA extracted with a Qiagen DNeasy Blood & Tissue Kit. DNA concentration was determined on a Nanodrop ND-1000 spectrophotometer.

To measure relative Wolbachia load, we performed two quantitative PCR (qPCR) assays. The first targeted Dfd, a single-copy nuclear gene in D. melanogaster (Dfd For 5' GTAGCGAAGAAACCCACCAA 3'; Dfd Rev 5' ACGCTCCACTCACCTCATTC 3'). The second used the primers wspFQALL and wspRQALL to target the wsp gene of Wolbachia (Osborne et al. 2009). Each 10 ml reaction contained 10 mM Tris 8.0, 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.25 mM SYBR green, 5% DMSO, 0.25 mM of each primer, Taq Polymerase, and 25 ng of DNA. Reactions were run in triplicate on a Roche LightClycler 480 with the following conditions: one cycle of 95°C for 5 min, followed by 45 cycles of 95°C for 15 sec, 60°C for 30 sec and 72°C for 10 sec. For each fly line, we tested two pools of 20 flies that were sampled from separate vials. A known Wolbachia-free fly was used as a negative control. For each line, we calculated relative Wolbachia density as $2^{(CPDfd-CPwsp)}$.

To test whether there was population-level variability in endosymbiont density, we ran an ANOVA on a Phylogenetic Generalized Least Squares (PGLS) model that tested for the

effect of population while controlling for the phylogenetic relationships between *w*Mel lineages. Phylogenetic correlations among *w*Mel strains were derived from the MrBayes analysis described above. We performed PGLS analyses using both Brownian Motion and Ornstein-Uhlenbeck Motion models. Analyses were conducted using the R packages ape v3.0-3 (Paradis et al. 2004) and nlme v3.1-102 (Pinheiro et al. 2012).

Results

Genome alignments and variant discovery

Wolbachia genome: Based on alignment to the W. pipientis *w*Mel reference genome, 65 of the 91 D. melanogaster lines showed strong evidence of Wolbachia infection (Table 2.1). For each of these lines, Mosaik mapped more than 90,000 reads to the *w*Mel reference genome, giving an average read depth of 7 or greater for each line (Early & Clark 2013). Conversely, 25 lines had fewer than 1,500 reads that mapped to the *w*Mel reference and so were considered Wolbachia-free. One line (B59) was intermediate to these two groups with 12,500 mapped reads. This could indicate an unusually low level of Wolbachia infection or a small amount of contamination. Because of the low genome coverage (about 1x), genotype calls could not be made with reasonable accuracy, and we excluded line B59 from the subsequent Wolbachia analyses.

Across the 65 Wolbachia-infected lines used in the subsequent analyses, GATK made base calls at 1,134,595 positions within the 1,267,782 bp genome (89.5%) and called single nucleotide polymorphisms (SNPs) at 174 positions. With additional filtering, we removed 22 sites where only heterozygous calls were made or where more than 10 lines were called as heterozygous or missing. Because of the repetitive nature of the *w*Mel genome, there was a

high probability that these calls resulted from misalignments, a hypothesis supported by the observation that 55% of these discarded calls were within 20 bp of a second low-confidence site. After this additional filtering, the final dataset contained 145 SNPs, of which 51 were detected in only a single line (Early & Clark 2013). Assuming each fly line carried a single *w*Mel copy (as discussed in the Methods), average genome-wide nucleotide diversity (π) across all populations was 1.8×10^{-5} (Table 2.2).

TABLE 2.2. Wolbachia and mitochondrial genomic diversity.

	N	π	*S*	*D*
***w*Mel**				
All	65	1.8×10^{-5}	145	-1.1
Beijing	9	2.8×10^{-5}	75	0.7
Ithaca, NY	14	4.0×10^{-6}	24	-1.7
Netherlands	12	3.1×10^{-6}	12	-0.5
Tasmania	14	3.7×10^{-6}	29	-2.3
Zimbabwe	16	7.9×10^{-6}	50	-1.7
mtDNA				
All	91	1.02×10^{-3}	147	-1.63
Beijing	16	1.51×10^{-3}	112	-1.48
Ithaca, NY	19	2.76×10^{-4}	33	-2.28
Netherlands	19	1.05×10^{-3}	73	-1.07
Tasmania	19	4.34×10^{-4}	26	-0.56
Zimbabwe	18	3.84×10^{-4}	39	-2.05

N, number of lines; π, average pairwise nucleotide diversity;
S, number of segregating sites; *D*, Tajima's D

Using our GATK pipeline and subsequent filtering, we identified 22 single bp indels within the *w*Mel genome. Pindel analysis identified six deletions and two small insertions (Early & Clark 2013). No inversions, tandem duplications, or IS-element insertions were identified. We compared all Pindel-identified breakpoints to those described in Riegler et al. (2005), and found no evidence that any of our samples deviated from the *w*Mel genotype. It is possible, however, that due to repetitive flanking sequences, this approach was unable to detect the inversion that differentiates *w*Mel from *w*Mel2 and *w*Mel3.

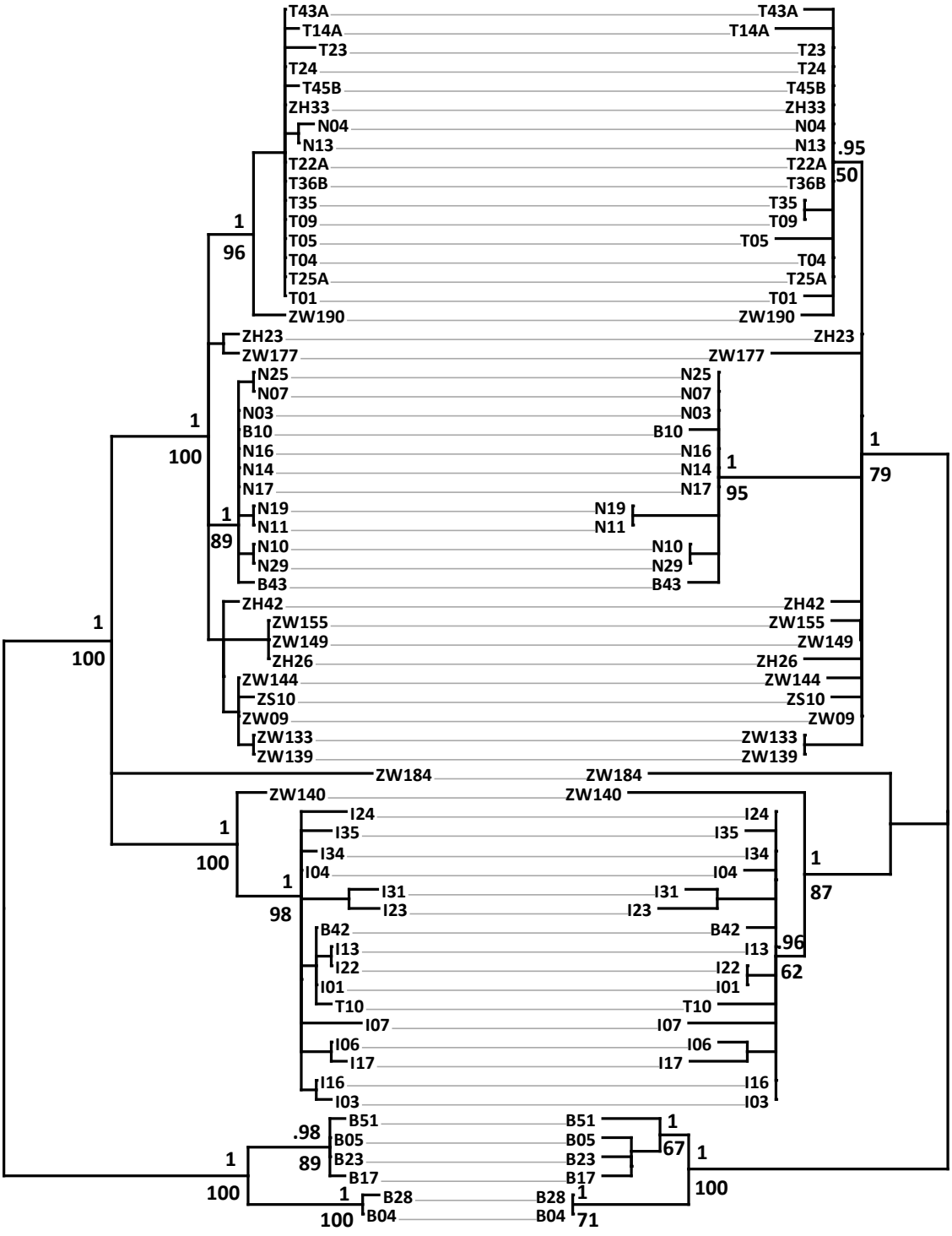
Mitochondrial genome: After MOSAIK alignment, the average mitochondrial read depth was 303.7 (Early & Clark 2013). For our analyses, we considered the GATK nucleotide calls for the first 14,916 bp of the chromosome. This includes all the coding regions, but excludes the repetitive AT-rich region where short-read alignments were unreliable. Within this region, calls were made at 14,661 positions (98.3%). GATK identified 166 SNPs, eight of which failed to pass our additional filters (Early & Clark 2013). Of the 158 SNPs in our final dataset, 11 were fixed within our sample (representing differences with the reference only) and 55 were singletons. In relation to the reference, Pindel analysis identified one 6 bp deletion present in all of our lines and one 5 bp insertion present in a subset of the lines (Appendix 2, Table S5). We found no inversions or tandem duplications. Average genome-wide nucleotide diversity (π) across all populations was 1.02×10^{-3} (Table 2.2).

Heteroplasmy: To determine whether any of our inbred fly lines were heteroplasmic, we closely examined all sites where multiple alleles were called within a single fly line. To be as rigorous as possible, we also analyzed 29 heterozygous *w*Mel sites that were filtered from our final dataset (making 40 sites in total for *w*Mel and seven for mtDNA). When haplotypes within a fly differed by only one or two SNPs, we could not determine whether heteroplasmy

was caused by horizontal transfer of a closely related haplotype or by a mutation within the maternal lineage. We therefore disregarded extremely low levels of heteroplasmy and instead determined whether flies carried two divergent mtDNA or *w*Mel haplotypes (for instance, haplotypes from both clades I and III, which are known to segregate in the same geographic areas). Upon examination, no fly lines had more than two heterozygous sites in their mtDNA (17 had one and a single line had two). While more diverse, the *w*Mel data yielded similar results. Examining only the sites in our final dataset, 12 lines had a single heterozygous *w*Mel site while two lines had two. Only a single fly line contained three heterozygous sites, but the alleles at these sites were not consistently shared with a one single known *w*Mel clade. The heterozygous calls in the low-confidence *w*Mel sites formed no informative pattern. In the wild, upwards of 14% of D. melanogaster may carry multiple mtDNA haplotypes (Nunes et al. 2013), however, our failure to find segregating divergent cytotypes here is not unexpected: these lines have been maintained in the lab beyond the 100 generations that Nunes et al. (2013) estimates is needed for complete sorting of mtDNA haplotypes. We conclude that the mtDNA heterozygosity seen is not due to paternal leakage of disparate haplotypes, but instead is some combination of within fly mutation and sequencing error. For this reason we sampled a single haplotype from each line for all other analyses.

Read depth comparison: We standardized mitochondrial read depth (as described in Methods) then compared standardized read depth in infected and uninfected lines and found no significant difference between the two groups (ANOVA, $P = 0.1233$). Similarly we tested whether Wolbachia read depth correlated with mitochondrial read depth and found no significant correlation (Pearson's product-moment correlation, $P = 0.7966$). These results are concordant with earlier findings in parasitic wasps (Mouton et al. 2009).

Figure 2.1. Tanglegram showing concordance of the A) wMel and B) mitochondrial phylogenetic trees.
Only infected lines are included in the mitochondrial tree. Trees are midpoint rooted. The first letter of the line name represents the geographic population of origin: B, Beijing, China; I, Ithaca, NY; N, Netherlands; T, Tasmania; Z, Zimbabwe.



A. *w*Mel

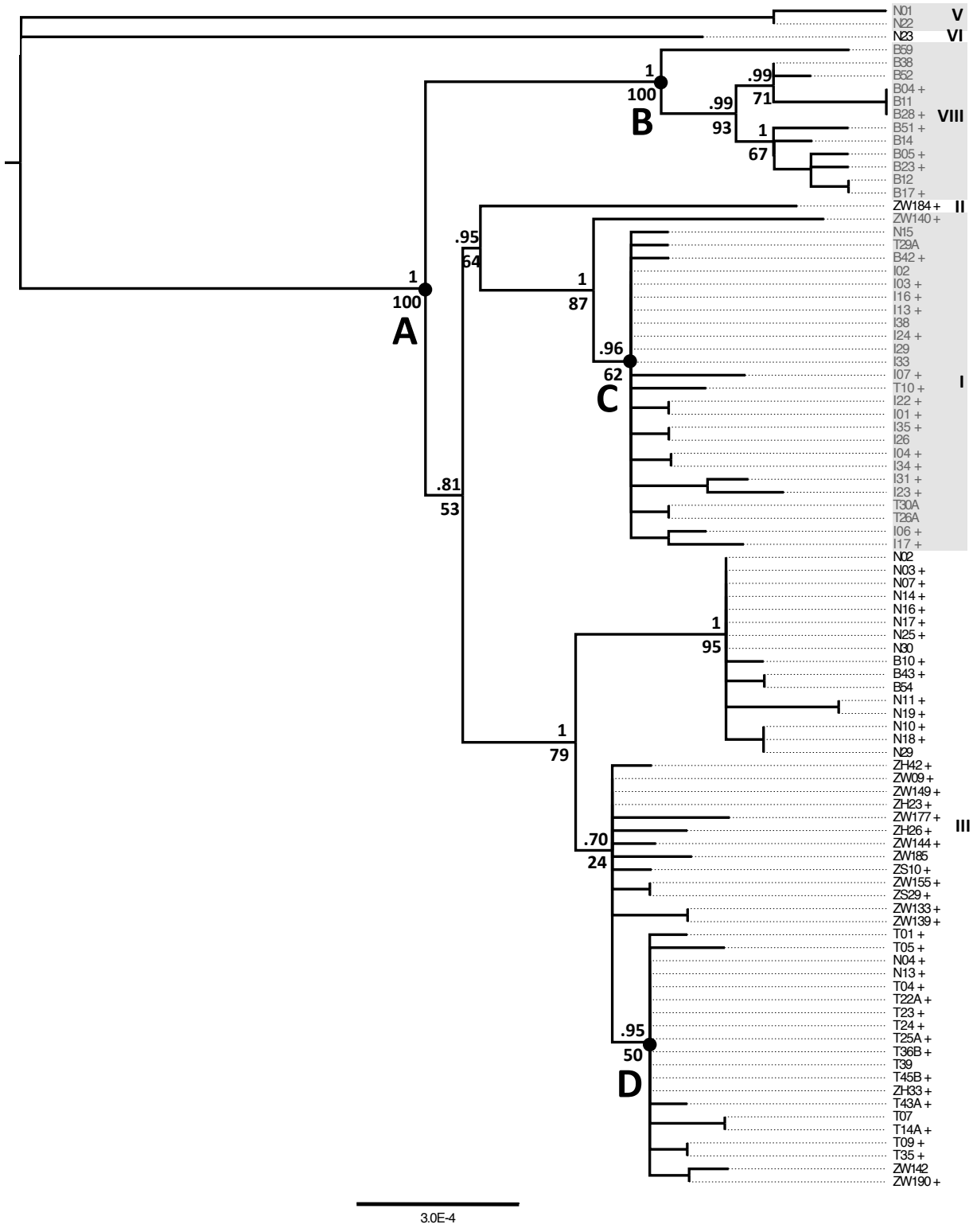
B. mtDNA

Lateral genetic transfer: Previously, transfer of Wolbachia genetic material into a host nuclear genome has been observed (Hotopp et al. 2007; Kondo et al. 2002; Nikoh et al. 2008). To test whether any portion of the *w*Mel genome has been transferred into the nucleus of any of our D. melanogaster genomes, we examined pairs of reads where one read mapped to the wMel genome while the paired read mapped to the D. melanogaster genome. For such paired reads, the only regions of the D. melanogaster genome mapped with an aligned read depth greater than two were highly repetitive and noninformative, providing no evidence of lateral genetic transfer in our sample.

Cytoplasmic haplotypes show geographic structuring

For both the *w*Mel and mtDNA data, we constructed phylogenetic trees with three methods: maximum likelihood using RAxML v7.2.8 (Stamatakis 2006), Bayesian inference using MrBayes v3.2 (Ronquist et al. 2012), and Bayesian inference using a strict molecular clock in BEAST v1.7.2 (Drummond & Rambaut 2007). The *w*Ri reference sequence or the D. simulans mitochondrial sequences described in the Methods were included to infer the root. Ignoring branching patterns with little support, all methods yielded identical topologies for the Wolbachia data (Figure 2.1A). For the mitochondria data, all methods resulted in identical branching patterns (again ignoring branches with little support), but these methods differed in the placement of the root. Because of this difficulty in resolving the root of the phylogeny, we constructed a phylogenetic network using the Neighbor-Net method in SplitsTree4 (Huson & Bryant 2006). The results showed the presence of conflicting phylogenetic signals in the mitochondria data (Appendix 2, Figure S1A). Because the haplotype divergence was so low, we did not have the power to test whether this pattern could have resulted from recombination.

Figure 2.2. Mitochondrial phylogenetic tree constructed with RAxML.
Major clades are marked on the right. Values above nodes are Bayesian probabilities calculated with MrBayes. Those below the nodes are bootstrap values calculated with RAxML. Lines followed by a + carry a Wolbachia infection. The ages of the marked nodes were calculated with a Bayesian Skyline analysis in BEAST using internal calibration points at nodes C and D. 95% HPD intervals are noted in parentheses. Root, 2,239 ya (1,100 – 3,592); A, 957 ya (462 – 1,556 ya); B, 425 ya (137 – 767 ya); C, 202 ya (91 – 311 ya); D, 192 ya (108 – 279 ya). The tree is midpoint rooted.


N01
N22
N23
V
VI
B59
B38
B52
B04 +
B11
B28 +
B51 +
B14
B05 +
B23 +
B12
B17 +
VIII
ZW184 +
ZW140 +
II
N15
T29A
B42 +
I02
I03 +
I16 +
I13 +
I38
I24 +
I29
I33
I07 +
T10 +
I22 +
I01 +
I35 +
I26
I04 +
I34 +
I31 +
I23 +
T30A
T26A
I06 +
I17 +
I
N02
N03 +
N07 +
N14 +
N16 +
N17 +
N25 +
N30
B10 +
B43 +
B54
N11 +
N19 +
N10 +
N18 +
N29
ZH42 +
ZW09 +
ZW149 +
ZH23 +
ZW177 +
ZH26 +
ZW144 +
ZW185
ZS10 +
ZW155 +
ZS29 +
ZW133 +
ZW139 +
III
T01 +
T05 +
N04 +
N13 +
T04 +
T22A +
T23 +
T24 +
T25A +
T36B +
T39
T45B +
ZH33 +
T43A +
T07
T14A +
T09 +
T35 +
ZW142
ZW190 +
1
100
B
.99
71
.99
93
1
67
1
100
A
.95
64
1
87
.96
62
C
.81
53
1
95
1
79
.70
24
.95
50
D
3.0E-4

The high mutation rate of D. melanogaster mtDNA, however, makes it likely that this pattern is the result of recurrent mutation in divergent lineages. We noted that one SNP and two indels from our dataset arose independently in a set of mutation accumulation lines (Haag-Liautard et al. 2008). In addition, 23 polymorphic sites within our lines are also known SNP locations in D. simulans mtDNA (Ballard 2000). No conflicting signals were detected in the Wolbachia data (Appendix 2, Figure S1B). The final rooting of the mtDNA tree is based on mid-point rooting and BEAST analyses, and is concordant with the rooting of the Wolbachia tree (Figure 2.1B).

Our sample contained five of seven previously defined mtDNA clades (Fig 2; Ilinsky 2013; Richardson et al. 2012). In addition, we describe here an additional clade (VIII), which segregates at a high frequency in the Beijing population and which has also been identified in flies carrying the *w*Mel2 haplotype (Chrostek et al. 2013). To further determine the extent to which our samples captured the full breadth of global D. melanogaster mitochondrial diversity, we constructed a mitochondrial tree that included both our samples and the 19 partial D. melanogaster mitochondrial genomes currently available in GenBank. In addition, we aligned short-read mtDNA sequences from whole-genome sequencing of a Canton-S line. These additional sequences clustered within or near already identified haplotypes showing that our samples capture a wide range of extant global genetic diversity (Appendix 2, Figure S2).

As in Richardson et al. (2012) and Ilinsky (2013), our cytoplasmic genomes show strong geographic structuring. All sampled populations contained one high frequency cytotype and with the exception of the Ithaca, NY population, at least one lower frequency cytotype (Figure 2.1 and 2.2; Table 2.1; Appendix 2, Figure S4). Despite this structuring, however, most cytotypes are not geographically isolated. The exceptions include clade VIII, which

was found only in the Beijing population, and clades V and VI, which were only detected in the Netherlands.

Evidence for strictly vertical wMel transmission with occasional loss events

If *w*Mel is transmitted exclusively maternally, it will show a tight evolutionary correlation with D. melanogaster mtDNA. Alternatively, if horizontal transmission has played a role in shaping *w*Mel evolution, we expect to see at least one of three possible patterns. First, we could directly find multiple wMel haplotypes within a single inbred fly line. Second, we could infer past co-infection by detecting recombination between divergent *w*Mel haplotypes. Or third, we could infer horizontal transmission or paternal leakage by finding the same *w*Mel haplotype associated with different mitochondrial backgrounds. As noted above, we found no support for the first two patterns: we did not detect multiple divergent haplotypes segregating within a single fly line, and the Neighbor-Net analysis provided no evidence of potential recombination. To test for the final pattern, we compared the branching patterns in the Wolbachia and mitochondrial phylogenies. With the exception of the Ithaca, NY population, all the sampled fly populations contained multiple segregating mitochondrial haplotypes, showing that opportunities exist for Wolbachia to contact new cytoplasmic backgrounds. However, like Richardson et al. (2012) we found the mtDNA and *w*Mel trees completely congruent, suggesting that horizontal transmission has not played a major role in recent *w*Mel evolution or ecology (Figure 2.1).

A second observation that comes from looking at infection patterns in the mtDNA phylogeny is that Wolbachia infections have been repeatedly lost since the most recent cytoplasmic coalescence (Fig 2; Richardson et al. 2012). All common mtDNA haplotypes con-

tained a mix of infected and uninfected cytoplasmic backgrounds, suggesting recent losses within these lineages. Within our sample, the two rare mitochondrial clades (V and VI) show no Wolbachia association, however, evidence from other studies suggest these lineages likely lost an ancestral infection. In Richardson et al. (2012), clade VI was shown to associate with Wolbachia. As for clade V, we noted that it clustered with the COI haplotype 10 whose members were infected with either *w*Mel or an undetermined strain of Wolbachia (Supporting Information, Fig S3; Nunes et al. 2008a). This provides evidence of past infections in both these lineages, and suggests that the coalescence of our mtDNA tree would also represent the MRCA of the extant global *w*Mel population. Most Recent Common Cytoplasmic Ancestor

Previous analyses have shown recent global shifts in the prevalence of particular *w*Mel haplotypes, as defined by five structural variants (Ilinsky & Zakharov 2007; Nunes *et al.* 2008b; Riegler *et al.* 2005). Specifically, since the 1960s the global frequency of *w*Mel-like haplotypes (to which all our *Wolbachia* samples belong) has risen sharply whereas the frequency of the *w*MelCS haplotype has declined rapidly. Our results support the hypothesis that this sweep was acting on standing variation (Richardson *et al.* 2012), as the *w*Mel haplotype and its corresponding mitochondrial background do not represent recent mutation events. To more precisely date the age of the major haplotypes and to determine the Most Recent Common Ancestor (MRCA) of all our cytoplasmic samples, we estimated divergence times and node ages of the mtDNA haplotypes with BEAST.

We conducted the analyses under two different sets of assumptions. The first model assumed a strict molecular clock, a constant population size, 10 fly generations per year (as in Richardson *et al.* 2012), and neutral evolution of third codon sites at the Haag-Liautard *et al.* (2008) estimate of the mitochondrial mutation rate. This gave a root age of 5,958 ya (95%

Highest Posterior Density (HPD): 4,216 – 7,886 ya) and internal node ages (defined in Figure 2.2) of: A, 2,578 ya (1,644 – 3,575 ya); B, 1,132 ya (482 – 1,843 ya); C, 620 ya (279 – 1,017 ya); and D, 568 ya (199 – 1,027 ya).

Our second set of models similarly assumed a strict molecular clock and neutral evolution at third codon positions. However, rather than calibrate the tree with the mitochondrial mutation rate, we placed age priors at nodes C and D (Figure 2.2), representing the estimated colonization of North America and Australia, respectively (approximately 200 years ago; David & Capy 1988; Keller 2007). Three separate models were run with different standard deviations placed on the clock-rate prior (6.2×10^{-7} mutations per site per year; standard deviations of 1×10^{-7}, 1×10^{-6}, and 1×10^{-5}). The node-calibrated models had higher marginal maximum likelihoods than the uncalibrated analysis, and the model with the greater support contained a clock standard deviation of 1×10^{-6} (Bayes Factor = 3.86, compared to the strict clock model). It dated the root of the tree to 2,239 years ago (95% HPD: 1100 – 3592 ya) and estimated the third codon position clock rate to be 1.75×10^{-6} substitutions per site per year (95% HPD: 9.5×10^{-7} – 2.6 x 10-6 substitutions per site per year). Internal node ages are given in Figure 2.2.

Molecular evolution of wMel

Because of its small effective population size, *Wolbachia* is expected to show reduced efficacy of selection. To test for this, we calculated the ratio of non-synonymous to synonymous amino acid substitutions (Ka/Ks) and compared SNP density as well as the estimated substitution rate within coding and non-coding regions. Calculated with the GY method in KaKs calculator (Zhang *et al.* 2006), average genome-wide Ka and Ks values were 1.58×10^{-5} and

1.89 x 10^{-5} respectively, giving a Ka/Ks ratio of 0.875 which is not statistically different from the expectation of neutral evolution with no constraint ($P = 0.508$), a finding also reached by Richardson *et al.* (2012). Additionally, we examined whether mutational patterns varied along the branches of the phylogenetic trees. By comparing mutations at the tips (singletons) to mutations on deeper branches, we found no significant differences in the ratio of non-synonymous to synonymous mutations.

To test whether SNPs were evenly distributed between coding and noncoding regions, we calculated the proportion of sites in both regions that were polymorphic. SNP density in intergenic regions (0.248 SNPs/kb) was higher than in protein-coding regions (0.112 SNPs/kb; Fisher's exact test (FET); $P = 2.34 \times 10^{-4}$; Table 2.3). Our BEAST models, however, did not find any significant difference in the substitution rate of coding versus intergenic regions (relative clock rates: Protein-coding, 0.928 (95% HPD: 0.075 – 2.35), Intergenic, 1.824 (95% HPD: 0.144 – 4.60)).

Table 2.3. Distribution of variant and invariant nucleotide sites in the wMel genomes.

	Ancestral State				Location		
	A	C	G	T	Protein-coding	Intergenic	Total
Invariant	366,331	199,828	198,234	370,074	961,702	133,078	1,134,450
Variant	30	36	37	25	108	33	145

Bacteria generally show a GC to AT mutational bias and maintain constant GC levels only through selection or after equilibrium nucleotide levels are reached (Hershberg & Petrov 2010). *w*Mel has an AT-rich genome (35% GC content), but it is unknown whether this is stable or whether the genome is evolving toward a still higher AT content. Consistent with

a GC to AT mutational bias, the results show a higher relative number of polymorphisms at ancestral C and G nucleotides compared to A and T nucleotides (FET, $P = 5.87 \times 10^{-7}$; Table 3). The majority of these mutations were transitions (Ti=102 and Tv=26), leading to an overall increase in AT content (70 GC-to-AT mutations versus 54 AT-to-GC mutations; FET, $P = 1.727 \times 10^{-6}$).

We did not find any evidence for codon selection. The *w*Mel genome shows strong codon usage bias that correlates with codon AT-content (Wu *et al.* 2004). In our dataset, the strength of codon bias did not correlate with the direction of synonymous codon mutations (Pearson's product-moment correlation, $P = 0.5291$), suggesting these observed patterns of variation are due to mutation, not selection.

BEAST analyses showed that the substitution rate in intergenic regions of the *w*Mel genome is 91 times slower than the substitution rate at third codon positions in the *D. melanogaster* mitochondrial genome. Assuming the Haag-Liautard *et al.* (2008) mitochondrial muta-

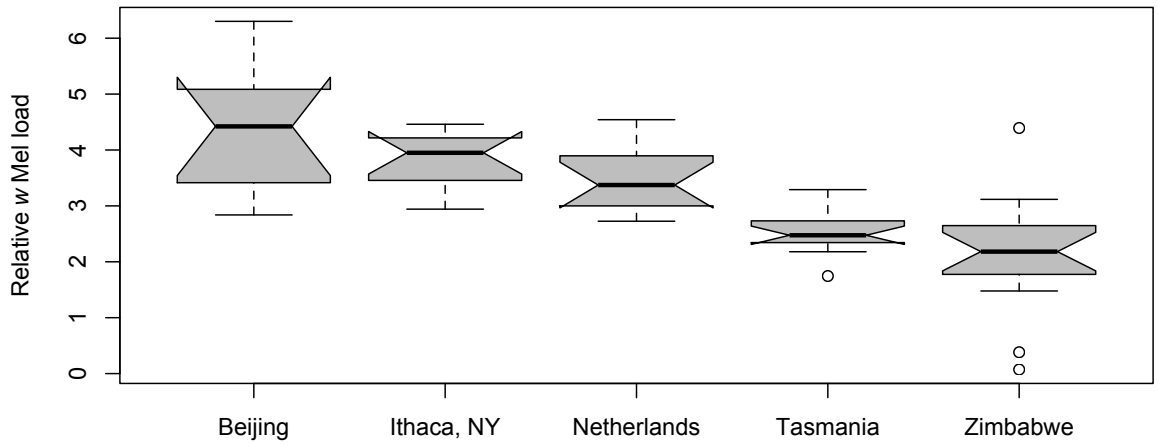


Figure 2.3. Average relative Wolbachia load within each population as determined by qPCR. In a PGLS model assuming Brownian Motion trait evolution based on the wMel phylogeny, the variation among populations is still significant (ANOVA, P = 0.0156).

tion rate, this yields a *w*Mel mutation rate of 6.8 x 10^{-10} substitutions per site per fly generation (95% HPD: 5.0 X 10^{-11} – 1.7 x 10^{-9}). This is almost identical to the estimate of 6.87 x 10^{-10} substitutions per site per generation calculated by Richardson *et al.* (2012).

Host effects on wMel within-fly density

Fitness effects conferred by *w*Mel on its fly host may vary with bacterial density (Osborne *et al.* 2009). We were therefore interested in examining whether populations varied in their *w*Mel load. Across our samples, depth of coverage of the *w*Mel genome was highly variable (Early & Clark 2013), an observation we validated with a qPCR analysis of additional replicates from each line. Despite the different methods, the qPCR and Illumina measurements had a Spearman's Rank Correlation of 0.787. The qPCR results confirmed that *Wolbachia* titre varies across lines and populations (Figure 2.3). Since our *w*Mel haplotypes are highly geographically structured, we constructed a Phylogenetic Generalized Least Squares (PGLS) model that tested the effect of fly population while controlling for the phylogenetic relationships between *w*Mel lineages. This model showed that fly populations significantly differ in their *Wolbachia* levels, a result that was robust under both a Brownian motion model (ANOVA, $P = 0.016$) and an Ornstein-Uhlenbeck model (ANOVA, $P = 0.0005$) of trait evolution.

Discussion

Because of the diversity of hosts it inhabits and the wide range of phenotypes it induces, the endosymbiont *Wolbachia* is a particularly intriguing study system. Until recently, phenotypic studies performed with *w*Mel have necessarily assumed genetic and phenotypic uniformity among infecting bacteria. Evidence over the last decade has shown that there is indeed genetic variation within this "clonal" infection (Ilinsky 2013; Ilinsky & Zakharov 2007; Nunes

et al. 2008b; Richardson *et al.* 2012; Riegler *et al.* 2005); however, many aspects of this variation, including its genomic extent, global distribution, and phenotypic effects, remain under-studied. Here, using a set of globally distributed populations, we combine an in-depth molecular genetic analysis with evidence of phenotypic variation among *w*Mel-infected host populations.

Global picture of wMel genomic diversity

Overall, our lines contain five of the seven previously described *D. melanogaster* cytotypes (Ilinsky 2013; Richardson *et al.* 2012). In addition, we name an additional mitochondrial and *w*Mel clade (VIII), which we found only in the Beijing population. The two known haplotypes that we do not recover currently have a low global frequency (clade IV; Richardson *et al.* 2012) or a limited geographic distribution (clade VII; Ilinsky 2013). While evidence suggests that all our cytoplasmic backgrounds are derived from a *Wolbachia*-infected ancestor, only four of our six cytoplasmic groups currently include *Wolbachia*-infected individuals and all of these represent *w*Mel-like infections. As in a previous study (Richardson *et al.* 2012), we find that genome-wide nucleotide diversity (π) among *w*Mel isolates is low (Table 2.2).

Concordant with past studies (Ilinsky 2013; Nunes *et al.* 2008a; Richardson *et al.* 2012), cytotypes display strong geographic structuring, with each population containing one major haplotype. A dominant haplotype could have arisen in each population due to drift, but it is likely that selection and local adaptation have played a role as gene flow is present among these populations. Phylogenetic analysis shows that the cytoplasmic associations between mitochondrial and *Wolbachia* haplotypes are stable and long-lived (Fig. 2.1; Richardson *et al.* 2012), so at this time, we can only hypothesize whether any selection has primarily acted

on mitochondria or *Wolbachia*. *Wolbachia* is, however, a likely target for selection. Recent studies have uncovered several fitness benefits that *w*Mel confers to its host including iron provisioning (Brownlie *et al.* 2009) and viral resistance (Teixeira *et al.* 2008). Indeed, some combination of selection and CI must allow for the maintenance of *Wolbachia* infections in natural populations. Otherwise, even the rare loss events seen in the wild (due to incomplete transmission from mother to offspring) would have resulted in a lower infection prevalence than what we, and others, have observed (Hoffmann *et al.* 1998; Ilinsky & Zakharov 2007; Richardson *et al.* 2012; Verspoor & Haddrill 2011).

Date of cytoplasmic coalescence

Linking known demographic information to our phylogenetic analysis, we date the cytoplasmic coalescence in *D. melanogatser* to approximately 2,239 ya (95% HPD: 1,100 – 3,592 ya). While overlapping with their 95% confidence intervals, this estimate differs from that recently proposed by Richardson *et al.* (2012; 8,008 ya, 95% BCI: 3,263-13,998 ya). Our decision to use a node-calibrated analysis arose from the observation that the major haplotypes in the two most recently founded populations (Ithaca, NY and Tasmania) displayed star-like topologies. As Richardson *et al.* (2012) proposed for a separate North American population, these haplotypes may have been repeatedly reintroduced to Tasmania and New York. A more parsimonious explanation for these star-like topologies, however, evokes a single founding event followed by a subsequent radiation within the population. Under this scenario, the ages of these clades should be no older than the colonization of the areas in question (approximately 200 years; David & Capy 1988; Keller 2007). Our uncalibrated coalescent analysis with a strict clock rate dates both of these nodes to approximately 600 ya, while the calibrated analysis

provides node estimates that are in line with demographic age estimates without deviating too far from our initial assumptions (Figure 2.2; Ithaca node: 202 ya, Tasmania node: 193 ya, third codon position clock rate: 1.75×10^{-6} substitutions per nucleotide per year).

Variation in determinants of w*Mel density*

Currently, little is known about the genetic interplay between *Wolbachia* and its hosts (Ikeya *et al.* 2009; Serbus *et al.* 2008; Yamada *et al.* 2011). Untangling these interactions will be key to understanding the evolution of these diverse symbioses. Here, we focus on one foundational phenotype that likely drives other phenotypic effects: *Wolbachia* titre. *Wolbachia* within-fly density has important implications for both partners as it correlates with levels of cytoplasmic incompatibility (Perrot-Minnot & Werren 1999; Poinsot *et al.* 1998; Unckless *et al.* 2009; Veneti *et al.* 2003) and potentially affects fitness benefits conferred to the fly (Osborne *et al.* 2012). Yet, despite its importance, we are only beginning to understand how bacterial density is regulated (Bordenstein *et al.* 2006; Serbus *et al.* 2011). Past studies have demonstrated a general effect of host genotypic variation on *Wolbachia* titer, but these studies have largely involved the transfer of *Wolbachia* infections among different host species (e.g., Bordenstein *et al.* 2003; McGraw *et al.* 2001).

We present evidence that *Wolbachia* titre varies among fly populations in a way that is independent of *w*Mel phylogeny. A recent analysis of *D. simulans* lines (Correa & Ballard 2012) showed that while *Wolbachia* ovarian density is highly variable in wild-caught females, this variability rapidly declines with laboratory rearing (within 19 generations). This observation suggests that the variation we see is not caused by the lingering effects of the environment, but is rather the result of intra-specific nuclear genetic variation among these differ-

ent populations. While we cannot conclude that the population-level variation reflects local adaptation and not the effects of drift, these results nevertheless point to the key role that host genotype plays in the regulation of *w*Mel density. Future studies could leverage the natural variation we describe here as a way of exploring further phenotypes and the specific genetic factors that mediate the interactions within this model symbiont-host system.

Acknowledgements

We wish to thank Margarida Cardoso Moreira, Rayna Bell and Robert Unckless for analytical advice and discussions, as well as Casey Bergman, Nancy Chen, Jae Young Choi, Vanessa Bauer DuMont, Jennifer Grenier, and three anonymous reviewers for helpful comments on this manuscript. This research was conducted using the resources of the Cornell Center for Advanced Computing and was supported by NIH grant R01 AI064950 to AGC and B. P. Lazzaro. This manuscript was originally published as: Early AE, Clark AG (2013) Monophyly of *Wolbachia pipientis* genomes within *Drosophila melanogaster*: geographic structuring, titre variation, and host effects across five populations. *Molecular Ecology,* 22: 5765-5778. It is reprinted here with permission.

References

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29: 2157-2167. doi: 10.1093/molbev/mss084

Baele G, Li WL, Drummond AJ, Suchard MA, Lemey P 2013. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Molecular Biology and Evolution 30: 239-243. doi: 10.1093/molbev/mss243

Baldo L, Bordenstein S, Wernegreen JJ, Werren JH 2006a. Widespread recombination throughout Wolbachia genomes. Molecular Biology and Evolution 23: 437-449. doi: 10.1093/Molbev/Msj049

Baldo L, Dunning Hotopp JC, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR, Hayashi C, Maiden MC, Tettelin H, Werren JH 2006b. Multilocus sequence typing system for the endosymbiont Wolbachia pipientis. Appl Environ Microbiol 72: 7098-7110. doi: 10.1128/AEM.00731-06

Ballard JWO 2000. Comparative genomics of mitochondrial DNA in Drosophila simulans. Journal of Molecular Evolution 51: 64-75.

Bordenstein SR, Marshall ML, Fry AJ, Kim U, Wernegreen JJ 2006. The tripartite associations between bacteriophage, Wolbachia, and arthropods. Plos Pathogens 2: 384-393. doi: 10.1371/Journal.Ppat.0020043

Bordenstein SR, Uy JJ, Werren JH 2003. Host genotype determines cytoplasmic incompatibility type in the haplodiploid genus nasonia. Genetics 164: 223-233.

Brownlie JC, Cass BN, Riegler M, Witsenburg JJ, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL 2009. Evidence for Metabolic Provisioning by a Common Invertebrate Endosymbiont, Wolbachia pipientis, during Periods of Nutritional Stress. Plos Pathogens 5. doi: 10.1371/Journal.Ppat.1000368

Chrostek E, Marialva MS, Esteves SS, Weinert LA, Martinez J, Jiggins FM, Teixeira L 2013. Wolbachia variants induce differential protection to viruses in Drosophila melanogaster: a phenotypic and phylogenomic analysis. Plos Genetics 9: e1003896. doi: 10.1371/journal.pgen.1003896

Correa CC, Ballard JW 2012. Wolbachia gonadal density in female and male Drosophila vary with laboratory adaptation and respond differently to physiological and environmental challenges. Journal of Invertebrate Pathology 111: 197-204. doi: 10.1016/j.jip.2012.08.003

David JR, Capy P 1988. Genetic-Variation of Drosophila-Melanogaster Natural-Populations. Trends in Genetics 4: 106-111. doi: 10.1016/0168-9525(88)90098-4

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491-498. doi: 10.1038/Ng.806

Drummond AJ, Rambaut A 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214. doi: 10.1186/1471-2148-7-214

Early AE, Clark AG 2013. Monophyly ofWolbachia pipientis genomes within Drosophila melanogaster: geographic structuring, titre variation, and host effects across five

populations. Molecular Ecology 22: 5765-5778. doi: 10.1111/mec.12530

Edgar RC 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797. doi: 10.1093/Nar/Gkh340

Friberg U, Miller PM, Stewart AD, Rice WR 2011. Mechanisms Promoting the Long-Term Persistence of a Wolbachia Infection in a Laboratory-Adapted Population of Drosophila melanogaster. Plos One 6. doi: 10.1371/journal.pone.0016448

Greenberg AJ, Hackett SR, Harshman LG, Clark AG 2010. A Hierarchical Bayesian Model for a Novel Sparse Partial Diallel Crossing Design. Genetics 185: 361-U551. doi: 10.1534/Genetics.110.115055

Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD 2008. Direct estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. Plos Biology 6: 1706-1714. doi: 10.1371/journal.pbio.0060204

Hershberg R, Petrov DA 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. Plos Genetics 6. doi: Doi 10.1371/Journal.Pgen.1001115

Hoffmann AA, Hercus M, Dagher H 1998. Population dynamics of the Wolbachia infection causing cytoplasmic incompatibility in Drosophila melanogaster. Genetics 148: 221-231.

Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MC, Giebel JD, Kumar N, Ishmael N, Wang SL, Ingram J, Nene RV, Shepard J, Tomkins J, Richards S, Spiro DJ, Ghedin E, Slatko BE, Tettelin H, Werren JH 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317: 1753-1756. doi: D10.1126/Science.1142490

Huson DH, Bryant D 2006. Application of phylogenetic networks in evolutionary studies.

Molecular Biology and Evolution 23: 254-267. doi: 10.1093/molbev/msj030

Huson DH, Scornavacca C 2012. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst Biol 61: 1061-1067. doi: 10.1093/Sysbio/Sys062

Ikeya T, Broughton S, Alic N, Grandison R, Partridge L 2009. The endosymbiont Wolbachia increases insulin/IGF-like signalling in Drosophila. Proceedings of the Royal Society B-Biological Sciences 276: 3799-3807. doi: 10.1098/Rspb.2009.0778

Ilinsky Y 2013. Coevolution of Drosophila melanogaster mtDNA and Wolbachia Genotypes. Plos One 8. doi: 10.1371/journal.pone.0054373

Ilinsky YY, Zakharov IK 2007. The endosymbiont Wolbachia in Eurasian populations of Drosophila melanogaster. Russian Journal of Genetics 43: 748-756. doi: 10.1134/S102279540707006x

Keller A 2007. Drosophila melanogaster's history as a human commensal. Current Biology 17: R77-R81. doi: 10.1016/J.Cub.2006.12.031

Klasson L, Westberg J, Sapountzis P, Nasiund K, Lutnaes Y, Darby AC, Veneti Z, Chen LM, Braig HR, Garrett R, Bourtzis K, Andersson SGE 2009. The mosaic genome structure of the Wolbachia *w*Ri strain infecting Drosophila simulans. Proceedings of the National Academy of Sciences of the United States of America 106: 5725-5730. doi: 10.1073/Pnas.0810753106

Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T 2002. Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proceedings of the National Academy of Sciences of the United States of America 99: 14280-14285. doi: 10.1073/Pnas.222228199

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,

Proc GPD 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079. doi: 10.1093/Bioinformatics/Btp352

McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-Loso T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG, Rumpho M, Sanders JG, Tautz D, Wernegreen JJ 2013. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci U S A 110: 3229-3236. doi: 10.1073/pnas.1218525110

McGraw EA, Merritt DJ, Droller JN, O'Neill SL 2001. Wolbachia-mediated sperm modification is dependent on the host genotype in Drosophila. Proceedings of the Royal Society B-Biological Sciences 268: 2565-2570. doi: 10.1098/Rspb.2001.1839

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303. doi: 10.1101/gr.107524.110

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069-2070. doi: 10.1093/bioinformatics/btq330

Medina M, Sachs JL 2010. Symbiont genomics, our new tangled bank. Genomics 95: 129-137. doi: 10.1016/J.Ygeno.2009.12.004

Moran NA, McCutcheon JP, Nakabachi A 2008. Genomics and Evolution of Heritable Bacterial Symbionts. Annual Review of Genetics 42: 165-190. doi: 10.1146/Annurev.Genet.41.110306.130119

Moran NA, McLaughlin HJ, Sorek R 2009. The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria. Science 323: 379-382. doi: 10.1126/Science.1167140

Mouton L, Henri H, Fleury F 2009. Interactions between Coexisting Intracellular Genomes: Mitochondrial Density and Wolbachia Infection. Applied and Environmental Microbiology 75: 1916-1921. doi: 10.1128/Aem.02677-08

Moya A, Pereto J, Gil R, Latorre A 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nature Reviews Genetics 9: 218-229. doi: 10.1038/Nrg2319

Nikoh N, Tanaka K, Shibata F, Kondo N, Hizume M, Shimada M, Fukatsu T 2008. Wolbachia genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes. Genome Research 18: 272-280. doi: 10.1101/Gr.7144908

Nunes MDS, Dolezal M, Schlotterer C 2013. Extensive paternal mtDNA leakage in natural populations of Drosophila melanogaster. Molecular Ecology 22: 2106-2117. doi: 10.1111/Mec.12256

Nunes MDS, Neumeier H, Schlotterer C 2008a. Contrasting patterns of natural variation in global Drosophila melanogaster populations. Molecular Ecology 17: 4470-4479. doi: 10.1111/J.1365-294x.2008.03944.X

Nunes MDS, Nolte V, Schlotterer C 2008b. Nonrandom Wolbachia Infection Status of Drosophila melanogaster Strains with Different mtDNA Haplotypes. Molecular Biology and Evolution 25: 2493-2498. doi: 10.1093/Molbev/Msn199

Osborne SE, Iturbe-Ormaetxe I, Brownlie JC, O'Neill SL, Johnson KN 2012. Antiviral Protection and the Importance of Wolbachia Density and Tissue Tropism in Drosophila

simulans. Applied and Environmental Microbiology 78: 6922-6929. doi: 10.1128/Aem.01727-12

Osborne SE, San Leong Y, O'Neill SL, Johnson KN 2009. Variation in antiviral protection mediated by different Wolbachia strains in Drosophila simulans. Plos Pathogens 5: e1000656.

Paradis E, Claude J, Strimmer K 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289-290. doi: 10.1093/Bioinformatics/Btg412

Perrot-Minnot MJ, Werren JH 1999. Wolbachia infection and incompatibility dynamics in experimental selection lines. Journal of Evolutionary Biology 12: 272-282.

Pinheiro J, Bates D, DebRoy S, Sarkar D, the R Development Core Team. 2012. nlme: Linear and Nonlinear Mixed Effects Models.

Poinsot D, Bourtzis K, Markakis G, Savakis C, Mercot H 1998. Wolbachia transfer from Drosophila melanogaster into D-simulans: Host effect and cytoplasmic incompatibility relationships. Genetics 150: 227-237.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reynolds KT, Hoffmann AA 2002. Male age, host effects and the weak expression or nonexpression of cytoplasmic incompatibility in Drosophila strains infected by maternally transmitted Wolbachia. Genetical Research 80: 79-87. doi: 10.1017/S0016672302005827

Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, Bergman CM 2012. Population Genomics of the Wolbachia Endosymbiont in Drosophila melanogaster. Plos Genetics 8: e1003129. doi: 10.1371/journal.pgen.1003129

Riegler M, Sidhu M, Miller WJ, O'Neill SL 2005. Evidence for a global Wolbachia replacement in Drosophila melanogaster. Current Biology 15: 1428-1433. doi: 10.1016/J.Cub.2005.06.069

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539-542. doi: 10.1093/sysbio/sys029

Serbus LR, Casper-Lindley C, Landmann F, Sullivan W 2008. The Genetics and Cell Biology of Wolbachia-Host Interactions. Annual Review of Genetics 42: 683-707. doi: 10.1146/Annurev.Genet.41.110306.130354

Serbus LR, Ferreccio A, Zhukova M, McMorris CL, Kiseleva E, Sullivan W 2011. A feedback loop between Wolbachia and the Drosophila gurken mRNP complex influences Wolbachia titer. J Cell Sci 124: 4299-4308. doi: 10.1242/jcs.092510

Soshnev AA, He B, Baxley RM, Jiang N, Hart CM, Tan K, Geyer PK 2012. Genome-wide studies of the multi-zinc finger Drosophila Suppressor of Hairy-wing protein in the ovary. Nucleic Acids Res 40: 5415-5431. doi: 10.1093/nar/gks225

Stamatakis A 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690. doi: 10.1093/bioinformatics/btl446

Taylor MJ, Bandi C, Hoerauf A 2005. Wolbachia bacterial endosymbionts of filarial nematodes. Adv Parasitol 60: 245-284. doi: 10.1016/S0065-308X(05)60004-8

Teixeira L, Ferreira A, Ashburner M 2008. The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. Plos Biology 6: 2753-

2763. doi: 10.1371/Journal.Pbio.1000002

Unckless RL, Boelio LM, Herren JK, Jaenike J 2009. Wolbachia as populations within individual insects: causes and consequences of density variation in natural populations. Proceedings of the Royal Society B-Biological Sciences 276: 2805-2811. doi: 10.1098/Rspb.2009.0287

Veneti Z, Clark ME, Zabalou S, Karr TL, Savakis C, Bourtzis K 2003. Cytoplasmic incompatibility and sperm cyst infection in different Drosophila-Wolbachia associations. Genetics 164: 545-552.

Verspoor RL, Haddrill PR 2011. Genetic Diversity, Population Structure and Wolbachia Infection Status in a Worldwide Sample of Drosophila melanogaster and D. simulans Populations. Plos One 6. doi: 10.1371/journal.pone.0026318

Werren JH, Zhang W, Guo LR 1995. Evolution and Phylogeny of Wolbachia - Reproductive Parasites of Arthropods. Proceedings of the Royal Society B-Biological Sciences 261: 55-63. doi: 10.1098/Rspb.1995.0117

Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadinejad N, Wiegand C, Madupu R, Beanan MJ, Brinkac LM, Daugherty SC, Durkin AS, Kolonay JF, Nelson WC, Mohamoud Y, Lee P, Berry K, Young MB, Utterback T, Weidman J, Nierman WC, Paulsen IT, Nelson KE, Tettelin H, O'Neill SL, Eisen JA 2004. Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: A streamlined genome overrun by mobile genetic elements. Plos Biology 2: 327-341. doi: 10.1371/Journal.Pbio.0020069

Yamada R, Iturbe-Ormaetxe I, Brownlie JC, O'Neill SL 2011. Functional test of the influence of Wolbachia genes on cytoplasmic incompatibility expression in Drosophila

melanogaster. Insect Molecular Biology 20: 75-85. doi: 10.1111/J.1365-2583.2010.01042.X

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871. doi: 10.1093/bioinformatics/btp394

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4: 259-263. doi: 10.1016/S1672-0229(07)60007-2

Zug R, Hammerstein P 2012. Still a Host of Hosts for Wolbachia: Analysis of Recent Data Suggests That 40% of Terrestrial Arthropod Species Are Infected. Plos One 7. doi: 10.1371/journal.pone.0038544

Chapter 3

Signatures of local adaptation in genes of the *Drosophila* immune system

Abstract

Life history traits that have been studied in *Drosophila* are highly polygenic, and yet we know little about how these traits adapt to novel environments. Here we explore complex trait adaptation by studying immune gene diversity and divergence in five populations of *Drosophila melanogaster*. This large-scale, population-based approach provides an important—and hitherto missing—counterpart to studies of long-term evolution. Our approach uses a set of carefully matched control genes to account for the effects of demography and recombination rate, allowing us to identify immune genes that have experienced stronger selection than non-immune genes. In addition, we examine discrete immune pathways to determine whether there is evidence for polygenic selection on particular immune functions. The idea of contrasting patterns of polymorphism across geographic populations is motivated by the idea that pathogens are likely non-uniform in distribution. We find that genes involved in virus and parasitoid wasp defense have experienced more rapid recent adaptation and display greater levels of population differentiation between ancestral and derived populations. Phagocytosis receptors are more highly differentiated between populations while effector genes show increased population differentiation in only one population pair despite being highly diverse. These genetic observations parallel known patterns of *Drosophila* pathogen biology and distributions, providing novel insight into the genetic architecture underlying phenotypic variation in immune competence. In addition, our results point to congruencies between temporal

and spatial variation in selection pressures and suggest that directional selection on immune genes has largely occurred as flies have encountered new environments.

Introduction

A current challenge in evolutionary biology is to elucidate the genetic architecture of local adaptation (Stapley, et al. 2010). In recent years, studies have addressed this question by examining genome-wide changes in allele frequencies across clines or between populations (Hubner, et al. 2013; Lamichhaney, et al. 2012; Pespeni, et al. 2012; Stapley, et al. 2010). In limited instances, these genome-wide patterns can even be connected back to specific phenotypic traits or environmental variables, greatly increasing our understanding of the genetic processes underlying phenotypic evolution (Jones, et al. 2012; Turner, et al. 2010).

While analytic approaches differ, these studies often rely on the detection of outlier loci that display patterns of high population differentiation or other extreme signatures of local selection. Such approaches are adept at detecting loci that have experienced strong directional selection, but they lack to power to identify small-effect loci, whose detection may be hindered by smaller selection coefficients and continued gene flow. Nevertheless, experimental evolution and quantitative genetic studies have repeatedly shown that small-effect loci contribute heavily to phenotypic traits, so their potential role in local adaptation should not be overlooked (Gibson 2012; Rockman 2012). To date, however, few studies have experimentally investigated these different evolutionary routes to determine the genetic architecture underlying adaptation in complex traits(Savolainen, et al. 2013; Scheinfeldt and Tishkoff 2013). (Savolainen, et al. 2013; Scheinfeldt and Tishkoff 2013)

Polygenic adaptation can be studied by grouping together genes based on their functions. This provides greater power and allows us to determine whether small signatures of selection are preferentially found in genes that contribute to a specific process. This "bottom up" approach first defines sets of genes known to be involved in the traits of interest. While many organisms lack such resources, we fortunately possess extensive knowledge of the genes affecting numerous *Drosophila melanogaster* phenotypes, allowing us to conduct such a study in this key model organism. Here, we have chosen to focus on a single ecologically relevant trait that has received particular attention from the research community in recent decades: immune defense.

Immune defense is a prime example of a complex trait subject to variation in local selection pressures. Immune genes are often evolutionary outliers, displaying fast rates of evolution and high population differentiation across multiple taxa including humans (Daub, et al. 2013; Fumagalli, et al. 2011; Quintana-Murci and Clark 2013), *Daphnia* (McTaggart, et al. 2012), mosquitoes (Crawford, et al. 2010; Waterhouse, et al. 2007), and bees (Chavez-Galarza, et al. 2013; Erler, et al. 2014). *Drosophila* species are no exception to this pattern and their immune genes are known to evolve faster than the genome average (Sackton, et al. 2007). Similarly, both genome-wide studies and studies of individual genes have highlighted examples of immune genes displaying unusually high population differentiation across *D. melanogaster* populations (Fabian, et al. 2012; Hubner, et al. 2013; Juneja and Lazzaro 2010). But while these observations suggest that *Drosophila* immune genes may be subject to unusually strong spatially variable selection, they provide only a partial picture of how immunity, as a whole complex phenotypic trait, is evolving.

"Immune competence" is in reality a suite of numerous complex phenotypes that are impacted not only by pathogen pressures but also by environmental factors and genotype-by-environment interactions (Howick and Lazzaro 2014; Lazzaro, et al. 2008; Lazzaro and Little 2009; McKean, et al. 2008). While their resistance to similar types of pathogens may be weakly correlated (Lazzaro, et al. 2006), flies show no evidence of cross-resistance to distinct pathogen classes (Kraaijeveld, et al. 2012). Similarly, resistance is often decoupled from tolerance, highlighting the numerous physiological processes that influence host survival (Ayres, et al. 2008; Ayres and Schneider 2009). Further complicating the process of adaptation, trade-offs involving immune defense and other key life history traits are documented (Kraaijeveld, et al. 2001; McKean, et al. 2008; Ye, et al. 2009), as are behavioral traits that lie outside the canonical immune system (Babin, et al. 2014; Kacsoh, et al. 2013).

Underlying this phenotypic complexity is an equally complex network of interacting pathways that regulate various immune functions (reviewed in Ferrandon, et al. 2007). The most thoroughly studied are the Toll and IMD pathways, which together regulate the humoral response. Pattern recognition receptors in these pathways bind common bacterial and fungal membrane components, triggering the downstream production of anti-microbial peptides (AMPs) and other microbicidal compounds. A robust cellular response coordinates the activities of specialized hemocytes such as phagocytes, which engulf foreign particles or necrotic cells. In fly larvae, lamellocytes and crystal cells, participate in defense against parasitic wasps through the encapsulation and melanization of the deposited eggs. Finally, *D. melanogaster* possesses an antiviral defense that largely operates through RNA-interference but also involves the JAK-STAT and Toll pathways. These immune responses are joined by members

of the JNK pathway, which together with other genes, contribute to various aspects of immune tolerance and resistance by mediating tissue repair and wound closure.

Since particular classes of immune genes respond preferentially to specific types of parasites and pathogens, we can use patterns of genetic adaptation to infer which pathogen classes exert strong selective pressure. This approach has successfully identified viruses and parasitoid wasps as drivers of long-term evolution along the *D. melanogaster* lineage and within the *melanogaster* sub-group, specifically (Kolaczkowski, et al. 2011; Obbard, et al. 2006; Obbard, et al. 2009; Salazar-Jaramillo, et al. 2014). Suggesting that these selection pressures are ongoing, controlled infection experiments have shown that fly populations differ in their responses to both of these pathogens (Dupas, et al. 2009). We know little, however, about the global genetic variation underlying this phenotypic differentiation.

Here, by studying how immune genes differ across populations, we make inferences about the variation in pathogen and parasite pressure across *D. melanogaster*'s range. We compare known immune genes with genomic controls matched for size, genome location, and local recombination rate. With this approach, we not only identify single-gene targets of local selection but also detect signatures of polygenic selection within specific pathways and gene classes. By combining methods that target both single-gene and polygenic selection, we are able to infer the genetic architecture underlying adaptive events in these populations. In addition, we find that many of the patterns described in studies of *Drosophila* gene divergence are recapitulated at the population level, suggesting a certain degree of parity between spatial and temporal variation in pathogen-imposed selection pressure.

Methods

Gene and fly line selection

Through literature searches, we assembled a list of 375 genes with well-supported immune function, many of which have appeared in previous large-scale studies (Obbard, et al. 2009; Sackton, et al. 2007). When possible, we assigned each gene to the immune pathway(s) or process in which it functions (Table 3.1) and classified it based on the role it plays ("functional class," *e.g.* recognition, phagocytosis receptor, signaling, negative regulator, effector, anti-microbial peptide). Categories were not mutually exclusive and some were nested within larger umbrella categories (for instance, all Toll and IMD genes were also included in the hu-

Table 3.1. Immune gene groupings based on pathway and biological function.

Immune Process	Genes	Basic Function(s)
Cellular	131	Hemocyte-mediated responses to all classes of parasites and pathogens
Encapsulation	37	Initial recognition and coating of parasioid wasp eggs
Phagocytosis	47	Cellular uptake of bacteria, fungi and necrotic cells
Epithelial	30	Regulation of bacteria and fungi on epithelial surfaces, including the gut, trachea, and reproductive tract
Humoral	144	Recognition and elimination of bacteria and fungi in the hemolymph through the production of antimicrobial compounds
IMD	59	Humoral pathway that targets mainly gram-negative bacteria
Toll	62	Humoral pathway that targets mainly gram-positive bacteria and fungi. Also triggered during parasitoid wasp attack
JAK-STAT	27	Signaling pathway implicated in responses to viruses, control of hemocyte differentiation, and regulation of humoral response
JNK pathway & wound repair	44	Epithelial repair and cell growth
Melanization & PO production	34	Cell-mediated response that responds to wounding, parasitoid wasp eggs and microbes
ROS production	14	Production of reactive oxygen species. Especially key in epithelial immune regulation
Viral defense	32	Destruction of virus through RNA interference; elimination of infected cells

moral class.) The full list of genes, as well as their pathway and functional class assignments can be found in Appendix 3, Table S1.

For each immune gene, we identified four control genes that were matched for size, genome location and local recombination rate. For size and position, we required that control genes had a total length (including introns) within either 1,500 bp of or 0.5-2x the total immune gene length, and we preferentially chose genes within 100 kb of the immune gene. Using the *Drosophila melanogaster* Recombination Rate Calculator v 2.3 (Fiston-Lavier, et al. 2010), we obtained the estimated local recombination rate of each gene and further required that immunity and control genes had comparable recombination rates. In all instances, the control genes were within 1 cM/Mb according to the RRC recombination estimate and within 2.5 cM/Mb using the more fine-scale recombination estimates of (Comeron, et al. 2012). If immune genes were found near the boundaries of known segregating inversions, we ensured that the matched controls were similarly within or outside of the inversion. In the event that more than four control genes fulfilled these requirements for a particular immune gene, we randomly chose four. Because of the restrictions, 17 immune genes had fewer than four controls in the final data set.

We obtained information on nucleotide polymorphisms within these genes (including their 1.5-kb flanking regions) for each member of the *D. melanogaster* Global Diversity Lines. These are a set of 84 inbred lines from five populations (15 from Beijing, China; 19 from Ithaca, NY, USA; 19 from the Netherlands; 18 from Tasmania; and 13 Zimbabwe) that have been sequenced to an average depth of 12x (Grenier *et al*, in prep). We masked genomic regions with poor "callability" (as defined in Grenier, *et al.,* in prep) to ensure that our analyses excluded genomic regions with large amounts of missing data. Using the Ensembl

database Variant Effect Predictor script (BDGP 5.25 assembly, release 64), we annotated the putative effect of each SNP within our genes of interest (nonsynonymous coding, synonymous coding, intronic, splice site, 3'UTR, and 5' UTR).

Analysis of D. melanogaster-D. simulans divergence

Using FlyBase (v. FB2013_06) we determined which of our genes had known one-to-one orthologs in *D. simulans*. We then downloaded all transcripts for each gene from FlyBase and performed codon-based alignments for all *D. melanogaster-D. simulans* pairs using PRANK (Loytynoja and Goldman 2005). For each transcript pair, we used custom Perl scripts to count the number of substitutions in nonsynonymous and four-fold degenerate sites and to calculate the length of the aligned regions after removing all gaps. For each gene, we then identified the transcript pair with the longest aligned region and the fewest number of nonsynonymous substitutions. We used these transcripts to represent the gene in all downstream analyses. This was a conservative approach adopted to minimize alignment errors. Incorporating polymorphism data from the ancestral Zimbabwe population, we used DFE-alpha v. 2.13 (Eyre-Walker and Keightley 2009) to determine the proportion of adaptive substitutions (a) and the relative rate of adaptive substitutions (ω_a, calculated relative to the substitution rate at four-fold degenerate sites) within each pathway or functional class. In brief, this method uses a maximum likelihood method to infer the distribution of fitness effects of new mutations from the folded site frequency spectrum. We ran DFE-alpha using a two epoch model allowing variable mutation effect sizes and variable shape parameters for the gamma distribution. Gene classes with fewer than 10 genes were excluded from the analysis.

Population genetic analyses

Using custom Perl scripts, we calculated derived allele frequency, pairwise nucleotide diversity (π), and pairwise F_{ST} for each SNP. These same statistics, as well as Watterson's θ, Tajima's D, K_{ST}, Hudson's S_{nn}, Fu and Li's D, and Fay and Wu's H, were also calculated for each gene as a whole and within fixed 1-kb windows across each gene. Although inbred, all fly lines retained at least some residual heterozygosity. When a fly line was heterozygous for a given site, we randomly sampled a single allele to use in all analyses. In all calculations, we used only biallelic SNPs. For π calculations, we accounted for missing data by adjusting the sample size at each SNP. We calculated F_{ST} at each SNP according to Weir and Cockerham (1984), adjusting for missing data as we did with π. The numerator and denominator were averaged separately across regions to calculate F_{ST} within windows and genes. Negative F_{ST} estimates were declared to be zero.

For each gene, we downloaded transcript information (*i.e.*, coordinates for exons, introns, 5'UTR, and 3'UTR regions) from FlyBase. We then used custom Perl scripts to calculate the length of these regions and the number of synonymous and nonsynonymous sites in each transcript. This information was used to calculate π and F_{ST} across synonymous and nonsynonymous sites using two approaches. First, we chose the longest transcript of a given gene and used it to define coding regions. Second, we used the average values of all transcripts. The two approaches yielded comparable results.

Class enrichment in genomic outliers

To determine which genes showed the strongest signs of differentiation among populations, we examined the distribution of test statistics across all our immune genes. Due to the large

variation in gene length, we compared the distribution of test statistics calculated within 1-kb windows. We then tested whether certain classes of genes showed evidence of higher over-all values of π, θ, pairwise F_{ST}, and global K_{ST}. For each gene class, we counted the number of windows found in the upper 5% of windows for each statistic. To obtain *p*-values, we ran 100,000 permutations to generate a null distribution for the number of windows we would expect to see for each class. In both instances, we excluded windows where more than 500 nt were masked. We conducted separate analyses for the major autosomes (2L, 2R, 3L, and 3R) and the X chromosome, excluding chromosome 4 because it contained only three immune genes.

Single gene tests using genomic controls

The preceding outlier test did not account for the effect of local genetic environment, and so ignored the effects of factors like recombination rate or selection on nearby alleles. To further control for these factors, we leveraged genetic data from our set of matched control genes to more conservatively assess the probability of selection at each immune gene. Accordingly, for each immune gene, we used its set of matched control genes to establish a null distribution for each test statistic calculated at single SNPs (population-level π, pairwise F_{ST}, and global K_{ST}). Using a Mann-Whitney U test, we then ascertained whether the SNPs within our gene of interest deviated from the background distribution. In addition, we compared the unfolded site frequency spectrum (SFS) of SNPs within each immune gene to the combined SFS of its four matched control genes, performing a Kolmogorov-Smirnov test to determine whether the two distributions differed. Deviations from the control distributions indicate that the immune gene has followed a different evolutionary trajectory from the surrounding control genes. Both

neutral processes and selection could contribute to differences in SNP-level F_{ST} and K_{ST} distributions, but further information from gene-level statistics like Tajima's *D* can further inform inferences of selection.

We repeated all tests using only SNPs of a certain type (*e.g.*, coding, nonsynonymous variants). For each test, we applied the Bonferroni correction to the p-values to account for multiple testing across multiple genes (n=370). We excluded from the analysis any immune gene with fewer than four matched control genes. All statistical analyses were carried out in R (R Development Core Team 2011).

Pathway analyses using genomic controls

To test for the presence of polygenic selection, we determined whether the genes in a given pathway or functional class deviated, as a set, from control expectations. Here, like for the single gene comparisons, we used the control genes to establish a null distribution for each test statistic. We then compared the corresponding distribution in the set of immune genes and tested for deviations using a Mann-Whitney U test or a Kolmogorov-Smirnov test (for the SFS). This process was conducted for each test statistic calculated within 1-kb windows or at the level of individual SNPs (for the SFS). As in the enrichment test, we chose to use 1-kb windows in order to control for the effect that gene length has on our ability to detect small regions of selection. As all windows were compared en masse, we note that this approach does not imply parity among corresponding windows. Rather, we assume that, as a group, the windows spanning a gene accurately capture the evolutionary patterns across the length of that gene. To control for differences in coding content among genes, we also repeated the pathway analyses using only windows with 500 bp coding sequence. We performed multiple test cor-

rections on all p-values as described for the single gene tests above, correcting for multiple testing across multiple processes (n=13) or functions (n=7).

Results

Assembly and annotation of gene lists

By trawling databases and the *D. melanogaster* literature, we assembled a list of 375 genes with well-supported immune function. We annotated each gene according to the biological pathway or process in which it functions (Table 3.1) and the general role it plays (recognition, signaling, or effector). Categories were not mutually exclusive and some genes were given multiple assignments (Sup Mat).

To test for the presence of spatially variable selection within these genes, we acquired sequence data and SNP calls for each gene from 84 inbred *D. melanogaster* lines from five populations (the Global Diversity Lines; Grenier *et al*, in prep). We discarded five genes that had less than 1500 nt or 50% sequence coverage within these lines, leaving us with a final study set of 370 immune genes. For each gene, we then chose four control genes that were matched for size, location, and recombination rate. In the case of 17 immune genes, we were unable to identify four adequate control genes, leaving us with a final set of 1449 control genes. Sequence data for each control gene was similarly acquired for each of the 84 fly lines.

Patterns of divergence between D. melanogaster *and* D. simulans *differ among gene classes*

Previous cross-species comparisons have shown that certain classes of *Drosophila* immune genes evolve more rapidly than others (Obbard, et al. 2009; Sackton, et al. 2007). As our gene set is larger and more comprehensive than ones used in past analyses, we tested whether this pattern also holds true for our set. Within our full set of genes, there were 1,249 control genes

and 293 immune genes with one-to-one orthologs in *D. simulans*. For each of these genes, we calculated nucleotide divergence at nonsynonymous and four-fold degenerate sites. Combining these values with ancestral polymorphism data from the Zimbabwe population, we then estimated the proportion of adaptive substitutions (α) and the relative rate of adaptive substitutions (ω_a) for each gene class (Eyre-Walker and Keightley 2009).

Measurements of α and ω_a differed across functional and process-based groups (Figure 3.1), demonstrating that there is significant heterogeneity in long-term evolutionary patterns

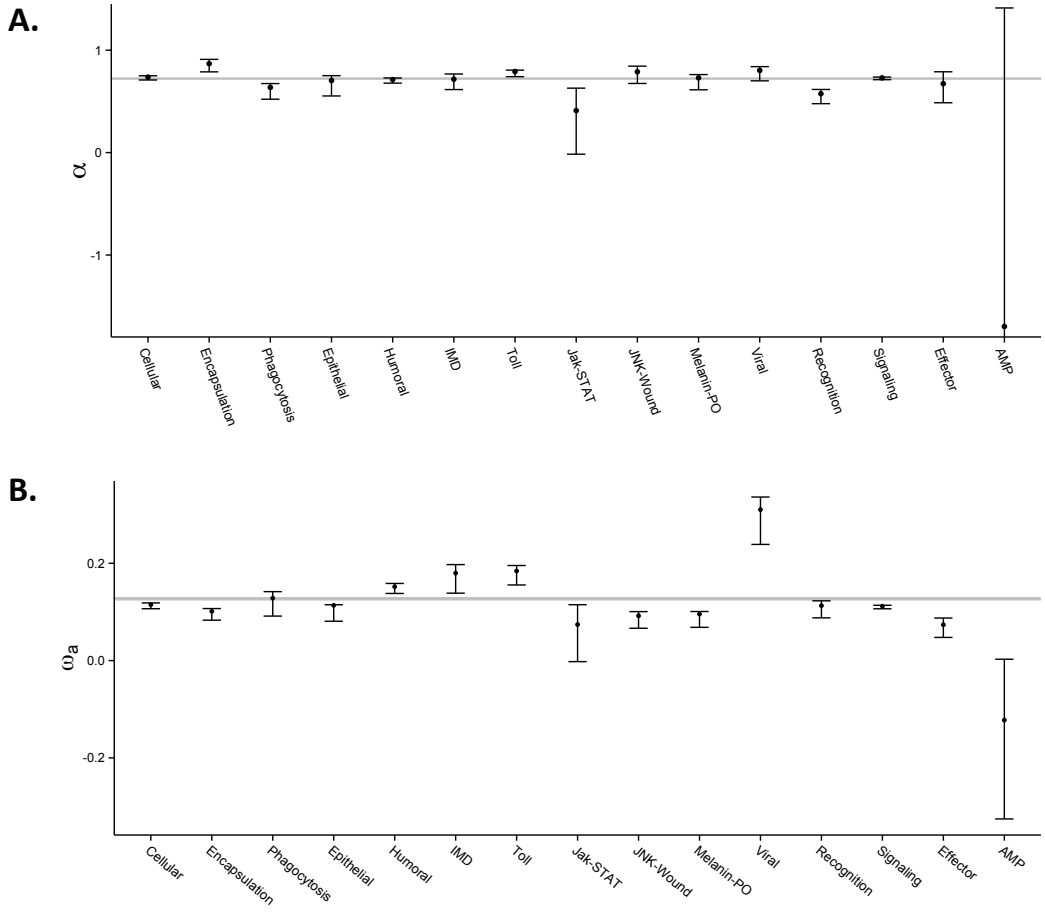


FIGURE 3.1. Adaptive divergence of Drosophila immune gene groups.
(A) Proportion of adaptive substitutions (α) and (B) relative rate of adaptive substitutions (ωa) calculated from D. melanogaster – D. simulans alignments using the DFE-alpha method of Keightley and Eyre-Walker (2009). Shown are the mean estimates with 95% CI calculated with jackknife resampling at the gene level. The grey line represents the mean and 95% CI as calculated with the set of 1,249 control genes. In figure A, the lower confidence interval for the group of AMP genes extends to -7.36.

across immune genes. The Toll pathway and encapsulation genes showed significantly elevated α relative to the control estimate (control: 0.7181 [0.7233,0.7266]; encapsulation: 0.8684 [0.7881, 0.9101]; Toll: 0.7916 [0.7413,0.8051]. Contrary to this, genes involved in phagocytosis, JAK-STAT signaling, and microbial recognition had α values that were low relative to the control set (Figure 3.1a; phagocytosis: 0.6367 [0.5211, 0.6740]; JAK-STAT: 0.4100 [-0.0158, 0.6291]; recognition: 0.5750 [0.4775, 0.6163]). The estimated α for our viral defense genes

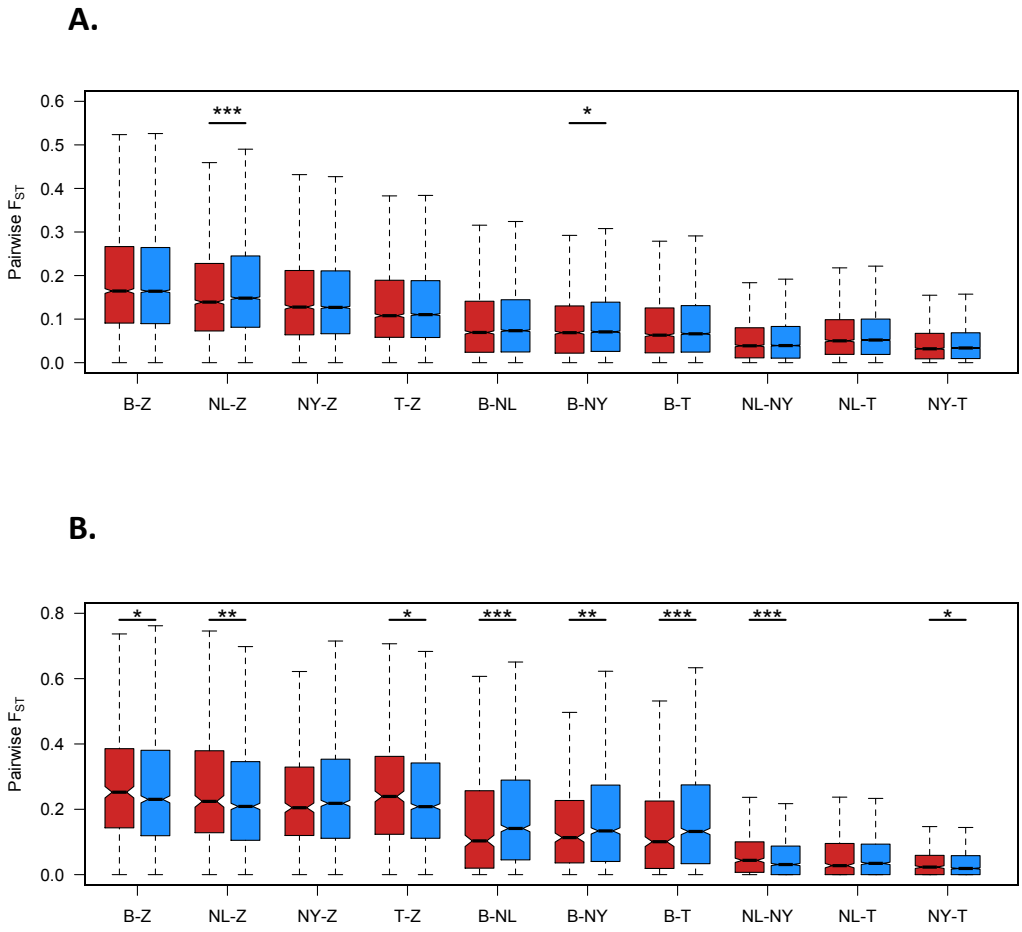


FIGURE 3.2. Patterns of pairwise FST on (A) the autosomes and (B) the X chromosome. Levels of FST reflect D. melanogaster's spread out of its ancestral range in sub-Saharan Africa (represented by the Zimbabwe population). Flies likely spread to Europe and Asia 10,000 years ago, only reaching Tasmania and North America through colonization by European populations within the last 200 years. Immune genes are in red and control genes in blue. Population samples are from Zimbabwe (Z), Beijing (B), the Netherlands (NL), New York (NY), and Tasmania (T). Differences between immune and control distributions were determined with one-sided Mann-Whitney U Tests. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

was elevated, but unlike Obbard, et al. (2009), the 95% CI overlapped with the control estimates, so the difference was not significant (0.8023 [0.7005, 0.8392]). These genes, however, did display a markedly higher ω_a relative to control estimates (Figure 3.1b; viral defense: 0.3102 [0.2387, 0.3361]; control: 0.1249 [0.1276,0.1289]). In addition, ω_a was elevated in

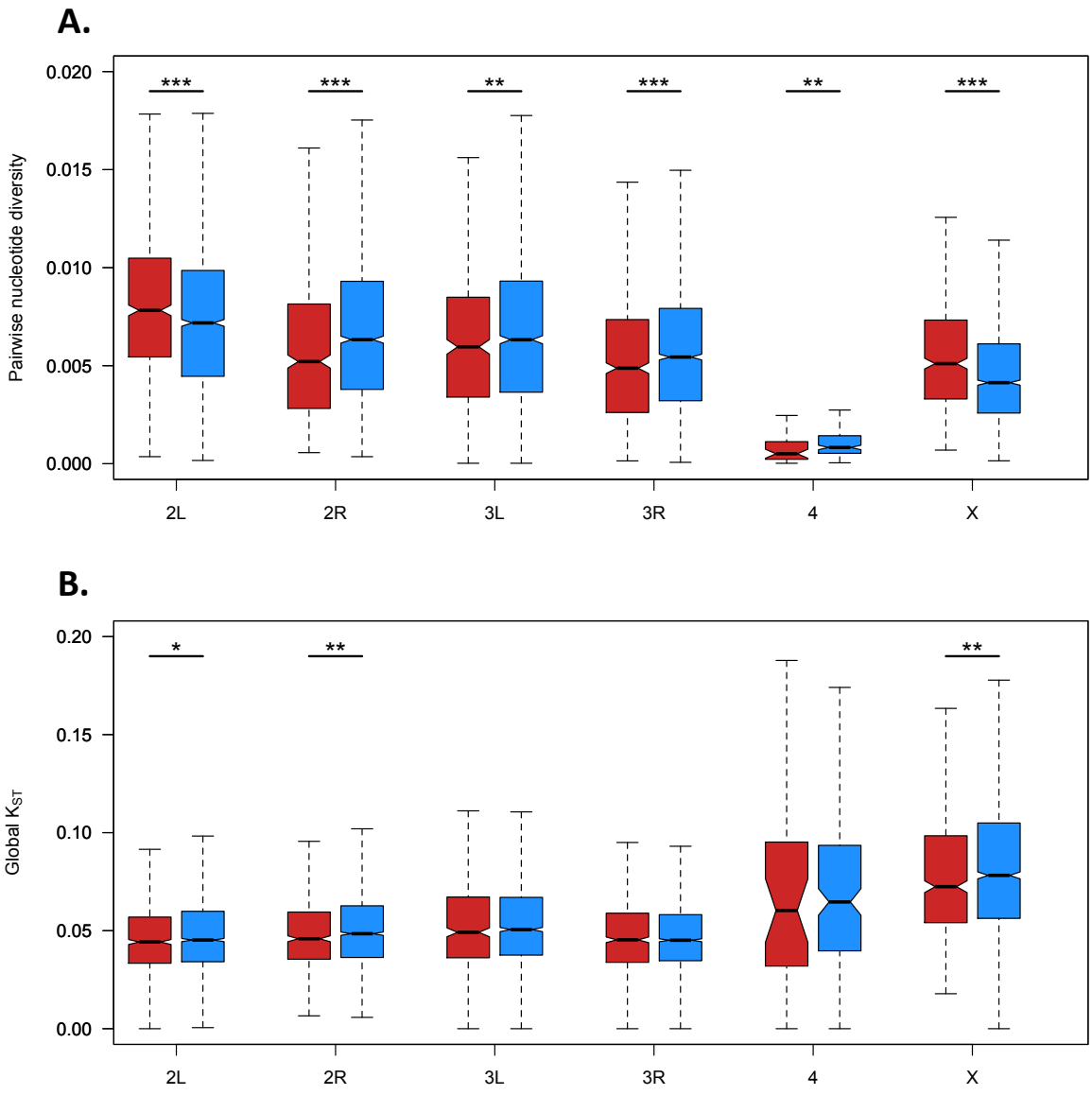


FIGURE 3.3. Global levels of nucleotide diversity and populations structure.
Patterns of (A) pairwise nucleotide diversity (π) and (B) global population structure (KST) differ across chromosomes and between immune and control genes. Immune genes are in red and control genes are in blue. Test statistics were calculated within fixed 1-kb windows that spanned the length of each gene. Differences between immune and control were determined with one-sided Mann-Whitney U Tests. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

humoral, IMD, and Toll genes (humoral: 0.1518 [0.1380, 0.1585]); IMD: 0.1797 [0.1385, 0.1972]; Toll: 0.1840 [0.1553, 0.1953]). In keeping with previous observations (Obbard, et al. 2009), AMPs showed no evidence of adaptive substitutions at the nucleotide level (α: -1.697 [-7.363, 1.412]; ω_a: -0.1223 [-0.3254, 0.00276]).

General patterns of nucleotide diversity and population differentiation

The general population genetic patterns within both the control and immune gene sets conformed to genome-wide observations made within these lines (Arguello, et al., in prep). For both sets of genes, F_{ST} between population pairs reflected known demographic patterns in *D. melanogaster*, which arose in sub-Saharan Africa, spread into Europe and Asia 10,000 years ago, and finally reached Australia and North America in the last 200 years (Figure 3.2; David and Capy 1988; Keller 2007; Laurent, et al. 2011; Thornton and Andolfatto 2006). Compared to genes on the four major autosomes (2L, 2R, 3L, and 3R), genes on the X chromosome harbor lower nucleotide diversity and higher F_{ST} and K_{ST} (Figure 3.3; Mann-Whitney U test for control and immune, $P < 0.00001$). Variation among autosomes also exists across all genes, highlighting the importance of local genomic effects in shaping patterns of variation. While these broad patterns hold true for both immune and non-immune genes, the two sets are not indistinguishable. Examining patterns of π in the 1-kb windows spanning our genes, we found that immune regions have significantly lower nucleotide diversity on chromosomes 2R, 3L, 3R, and 4, but significantly higher nucleotide diversity on chromosomes 2L and X (Figure 3.2; Mann-Whitney U test with $P < 0.01$ for all tests). In the individual populations, we observed similar trends for π, but they were not always statistically significant. Within 1-kb regions, global K_{ST} trended lower across immune regions on all chromosomes, although the

comparisons were only significant on 2L, 2R, and X (Figure 3.2; $P = 0.019$, $P = 0.00677$, and $P = 0.00886$, respectively). These patterns, however, were not detected when we compared measures of π and global K_{ST} at the full gene level.

Regions of high population structure are not uniformly distributed among gene classes

As shown above, patterns of *D. melanogaster-D. simulans* divergence vary among immune gene class, so we tested whether such variation also exists on the population level. To do so, we looked for signatures of local adaptation by calculating Tajima's *D*, Fay and Wu's *H*, Fu and Li's *D*, π, pairwise F_{ST}, and global K_{ST} for each immune gene.

We next looked at the distribution of these test statistics across all immune genes and determined whether genes of certain classes were significantly more likely to have extreme values. As the genes in our data set vary in size from 210 bp to 162 kbp, whole-gene comparisons would bias our results against detecting selected regions in large genes. We therefore calculated our summary statistics within fixed 1-kb windows that spanned the length of each gene. To test for an overrepresentation of a particular gene class among the extreme values, we identified windows in the upper (or lower for Tajima's *D*, Fay and Wu's *H*, Fu and Li's *D*) 5% and 1% of each test statistic's distribution, analyzing the X and autosomes separately. We determined the expected representation of each class within these tails through permutation tests that accounted for the length of each gene and the composition of each gene class. Using these permuted distributions, we established a significance threshold of $\alpha = 0.05$.

As with the patterns of long-term evolution, we found that signatures of recent adaptation are not uniformly distributed among gene classes. Effector genes, and in particular genes encoding antimicrobial peptides (AMPs), showed evidence of high nucleotide diversity

in all populations. In the absence of other evidence like significant Tajima's *D* or heightened population differentiation, this observation is consistent with the hypothesis that these genes are experiencing relaxed selective constraint, both globally and within individual populations. JAK-STAT genes showed signs of positive selection in Tasmania, where they were enriched in the lower tail of Tajima's *D* and Fu and Li's *D*, and showed heightened F_{ST} between Tasmania and Zimbabwe (as well as Ithaca and Zimbabwe). Of all the pathways, viral defense showed the strongest signs of extreme population structure. Viral defense was the only class of genes over-represented in the upper 1% and 5% of global K_{ST} values. Conversely, both humoral response genes and genes involved in ROS production were under represented. These patterns show that populations may have diverged more in their defenses against viruses than in their defenses against bacteria and fungi.

Comparisons of single immune genes to matched controls reveal selection candidates

Local genetic factors like recombination rate and background selection can influence our ability to detect selection. We therefore partially controlled for these effects by comparing the full set of SNPs in each of our immune genes to a distribution derived from the SNPs in four size- and position-matched control genes. The SNPs in the majority of the immune genes did not differ significantly from the background distributions established with control gene values. Still, after correcting for multiple testing over multiple genes (n=360), 60 immune genes showed elevated global K_{ST} or elevated F_{ST} in at least one pairwise comparison at $P \leq 0.05$ (Supplementary Table). Table 3.2 shows the 40 genes for which there is the strongest evidence of population differentiation. These genes showed elevated global K_{ST}, elevated F_{ST} in at least two pairwise comparisons, or elevated F_{ST} between a population and its most recent ancestral

population (*e.g.*, Ithaca and the Netherlands). Many of these genes also show other evidence of recent selection such as extreme values of Tajima's *D* or Fay and Wu's *H* (Table 3.2).

While no gene class was significantly over-represented in this group (Fisher's Exact Test), we still note that genes involved in melanization/PO production, encapsulation, and cellular recognition are particularly prominent. While the set of 40 genes represents 11% of our full gene set, 21% of melanization-PO genes, 16% of encapsulation genes, and 18% of cellular recognition genes were found within it.

Evidence for spatially variable, polygenic selection in viral and wasp defense genes

In addition to identifying large-effect candidate genes, we were interested in determining whether polygenic selection has played a role in shaping immune adaptation in these populations. To do this, we examined all genes within an immune group and looked for significant deviations from the distribution created with their combined control genes. By using these local genomic controls of unrelated function, we were largely able to account for patterns created by demography. Therefore, when we saw deviations between the immune genes and controls, we could more confidently attribute the differences to non-neutral forces acting on the immune set.

Within most pathways, pairwise F_{ST} and global K_{ST} levels were comparable for immune and control sets. When we did detect deviations, we tended to observe less population differentiation within immune genes than within controls. There were, however, two processes involved in parasitoid wasp defense that were notable exceptions to this trend: encapsulation and melanization/phenoloxidase production.

Table 3.2. Immune genes with significantly elevated global population structure or pairwise population differentiation.
Each listed gene displayed significantly elevated population differentiation for either global KST or pairwise FST when compared to its four control genes ($P < 0.05$ after Bonferroni correction; Population Structuring, High). In certain instances, these genes also showed significantly greater conservation between populations (Population Structuring, Low). As these genes are candidates for local adaptation, the populations with the lowest Tajima's D and Fay and Wu's H are listed along with the corresponding statistic value.

Gene		Function	Population Structuring	Lowest Fay & Wu's *H*	Lowest Tajima's *D*
Alk	FBgn0040505	Encapsulation; Melanization-PO; Signaling	High: Global K_{ST}, Z-B F_{ST}, Z-N F_{ST}, Z-I F_{ST} Low: B-I F_{ST}	B: -11.4916	B: -0.8882
armi	FBgn0041164	Viral (RNAi)	High: Global K_{ST}	I: -4.4036	Z: -0.3488
atilla	FBgn0032422	Encapsulation	High: B-NL F_{ST}, B-I F_{ST}, B-T F_{ST} Low: Z-I F_{ST}, Z-T F_{ST}	N: -5.3048	N: -0.8874
AttB	FBgn0041581	IMD; Effector	High: Global K_{ST}	N: -8.1732	I: -2.2631
CG6426	FBgn0034162	Humoral; Effector (lysozyme)	High: B-N F_{ST}, N-I F_{ST}, N-T F_{ST}	N: -3.2697	Z: -0.9951
CG8492	FBgn0035813	Humoral; Effector (lysozyme)	High: B-N F_{ST}, B-I F_{ST}	Z: -2.1229	Z: -1.1639
CHKov1	FBgn0045761	Viral	High: Global K_{ST}, Z-N F_{ST}, Z-I F_{ST}	N: -10.9416	N: -2.2074
CHKov2	FBgn0039328	Viral	High: Z-B F_{ST}, Z-N F_{ST}	N: -10.4638	N: -2.0493
daw	FBgn0031461	Toll; Melanization-PO; Signaling	High: Global K_{ST}, Z-N F_{ST}	B: -6.6928	B: 0.0488
dnr	FBgn0260866	IMD; Signaling	High: N-I F_{ST}, N-T F_{ST}	N: -8.9502	Z: -0.5597
drpr	FBgn0027594	Phagocytosis; Recognition	High: Z-N F_{ST}, N-T F_{ST} Low: B-N F_{ST}	N: -4.2667	Z: -0.6124
egh	FBgn0001404	Viral	High: N-I F_{ST}	I: -7.1174	Z: -0.6441
Gp150	FBgn0013272	Cellular; Recognition	High: Z-B F_{ST}	B: -8.5973	B: -1.2861
grh	FBgn0259211	Wound repair; Signaling	High: B-N F_{ST}, B-I F_{ST}, B-T F_{ST}	B: -15.2929	Z: -0.4743
Gr28b	FBgn0045495	Melanization-PO; Signaling	High: Z-B F_{ST}, Z-I F_{ST} Low: B-I F_{ST}	B: -11.0216	B: -1.0623
hep	FBgn0010303	JNK/wound repair; Melanization-PO; Signaling	High: Z-N F_{ST}, Z-I F_{ST}, N-T F_{ST} Low: N-I F_{ST}	N: -7.7362	N: -1.5296
ken	FBgn0011236	JAK-STAT; Signaling	High: B-I F_{ST}, B-T F_{ST}, N-I F_{ST}	B: -7.1112	B: -1.2046
LpR1	FBgn0066101	Humoral; Signaling	High: B-N F_{ST}, B-I F_{ST}	N: -5.6512	Z: -0.7475
LanA	FBgn0002526	Encapsulation; Recognition	High: N-T F_{ST}	Z: -5.7675	Z: -0.7211
lectin-24A	FBgn0040104	Cellular; Recognition (wasp)	High: Global K_{ST}	I: -3.0016	N: -1.0544

TABLE 3.2 (continued)

Gene		Function	Population Structuring	Lowest Fay & Wu's *H*	Lowest Tajima's *D*
Mekk1	FBgn0024329	General stress response; Signaling	High: Global K_{ST}, N-T F_{ST}	T: -3.7442	Z: -0.6687
NimC3	FBgn0001967	Phagocytosis; Recognition	High: Z-N F_{ST}, Z-I F_{ST}, Z-T F_{ST}	T: -5.4866	B: -1.8019
Nrg	FBgn0264975	Encapsulation; Recognition	High: N-T F_{ST} Low: B-I F_{ST}	T: -11.7725	Z: -0.4664
Ntf-2	FBgn0031145	Humoral; Signaling	High: N-I F_{ST}, I-T F_{ST} Low: Z-B F_{ST}, B-I F_{ST}	N: -5.8652	Z: -0.2835
NT1	FBgn0261526	Epithelial; Signaling	High: Z-N F_{ST} Low: B-I F_{ST}, B-T F_{ST}	N: -5.7324	N: -0.8989
pes	FBgn0031969	Phagocytosis; Recognition	High: Z-B F_{ST}, Z-N F_{ST}, Z-T F_{ST}	T: -10.8944	T: -0.9778
PGRP-LA	FBgn0035975	IMD; Epithelial	High: Global K_{ST}, N-T F_{ST} Low: B-N F_{ST}	N: -11.8994	N: -1.3265
proPO59	FBgn0261363	Encapsulation; Melanization-PO; Effector	High: B-N F_{ST}, B-I F_{ST}, B-T F_{ST}	N: -6.1747	B: -1.5227
puc	FBgn0243512	Wound repair (JNK)	High: N-T F_{ST}	N: -3.4266	B: -1.1195
RhoGEF3	FBgn0264707	Encapsulation; Signaling	High: N-I F_{ST} Low: B-T F_{ST}	I: -9.5710	I: -1.3104
Sr-CIII	FBgn0020376	Phagocytosis; Recognition	High: B-N F_{ST}, B-T F_{ST}	N: -3.2763	B: -0.5510
Stam	FBgn0027363	Cellular; Wound respair; Signaling	High: Z-B F_{ST}	B: -5.4089	B: -0.3274
Su(H)	FBgn0004837	Cellular; Melanization-PO; Signaling	High: Z-I F_{ST}, Z-T F_{ST} Low: B-T F_{ST}	T: -4.0428	T: -1.7064
Tab2	FBgn0086358	IMD; JNK; Signaling	High: N-T F_{ST}	B: -5.8242	Z: -0.3653
Takl2	FBgn0039015	Wound repair; Signaling	High: Z-N F_{ST}	N: -5.3509	N: -1.6589
Thor	FBgn0261560	Humoral; Phagocytosis; Signaling	High: Z-B F_{ST}	B: -2.3472	Z: -0.7586
Tl	FBgn0262473	Toll; Signaling	High: Z-N F_{ST}, Z-I F_{ST}, B-T F_{ST}, N-T F_{ST}, I-T F_{ST} Low: B-N F_{ST}	I: -8.4467	Z: -0.7070
Tsp68C	FBgn0043550	Cellular; Signaling	High: N-I F_{ST}, I-T F_{ST}	T: -4.3534	N: -0.0787
yellow-f	FBgn0041710	Melanization-PO; Effector	High: Z-B F_{ST}, B-T F_{ST}, N-I F_{ST}	B: -8.4771	B: -1.3559

After correction for multiple testing, we found that encapsulation genes had elevated global K_{ST} (Figure 3.4; $P < 0.00001$) and elevated K_{ST} in four population pairs: Beijing-Zimbabwe, the Netherlands-Zimbabwe, Tasmania-Zimbabwe, and the Netherlands-Ithaca (Mann-Whitney U test, $P \leq 0.05$). In addition, two populations (Zimbabwe and Tasmania) had significantly high S_{nn} values (Mann-Whitney U test, $P \leq 0.005$). Conversely, population differentiation was significantly reduced between Beijing and the other derived populations (pairwise K_{ST} and F_{ST}; Mann-Whitney U test, $P \leq 0.05$), suggesting that the high global K_{ST} is mainly driven by differences between the ancestral African populations and the derived populations. The patterns for melanization-PO genes were similar but less extreme. In our raw analysis, all pairwise combinations with Zimbabwe showed elevated F_{ST} and K_{ST}, but only two of the comparisons remained significant after correction for multiple testing (F_{ST} for Beijing-Zimbabwe and the Netherlands-Zimbabwe; Mann-Whitney U test, $P < 0.05$).

Complementing these signatures of high population differentiation, we found further evidence for selection on encapsulation genes within the derived populations. Within Beijing, the Netherlands, Ithaca, and Tasmania, the unfolded site frequency spectra (SFS) were significantly skewed to the right (Figure 3.5; K-S Test, $P < 0.05$). This pattern corresponds to a relative increase in high frequency derived alleles, a common hallmark of directional selection. A similar rightward skew was also present when we examined alleles found only in a single population (private alleles). In all populations except Tasmania, the unfolded SFS of private encapsulation alleles was shifted towards the right, although after a Bonferroni correction, this pattern only remained significant for the Netherlands and Zimbabwe (K-S test; $P < 0.0001$). Additionally, the distributions of both Tajima's D and Fay and Wu's H were lower in all four derived populations, although after correction for multiple testing, this shift was only signifi-

Figure 3.4. Population structure (global KST) within each immune gene class.
KST values were calculated within 1-kb windows for both immune (red) and control (blue) genes. Differences between immune and control were determined with one-sided Mann-Whitney U Tests. * P <0.05; ** P < 0.01; *** P < 0.001.

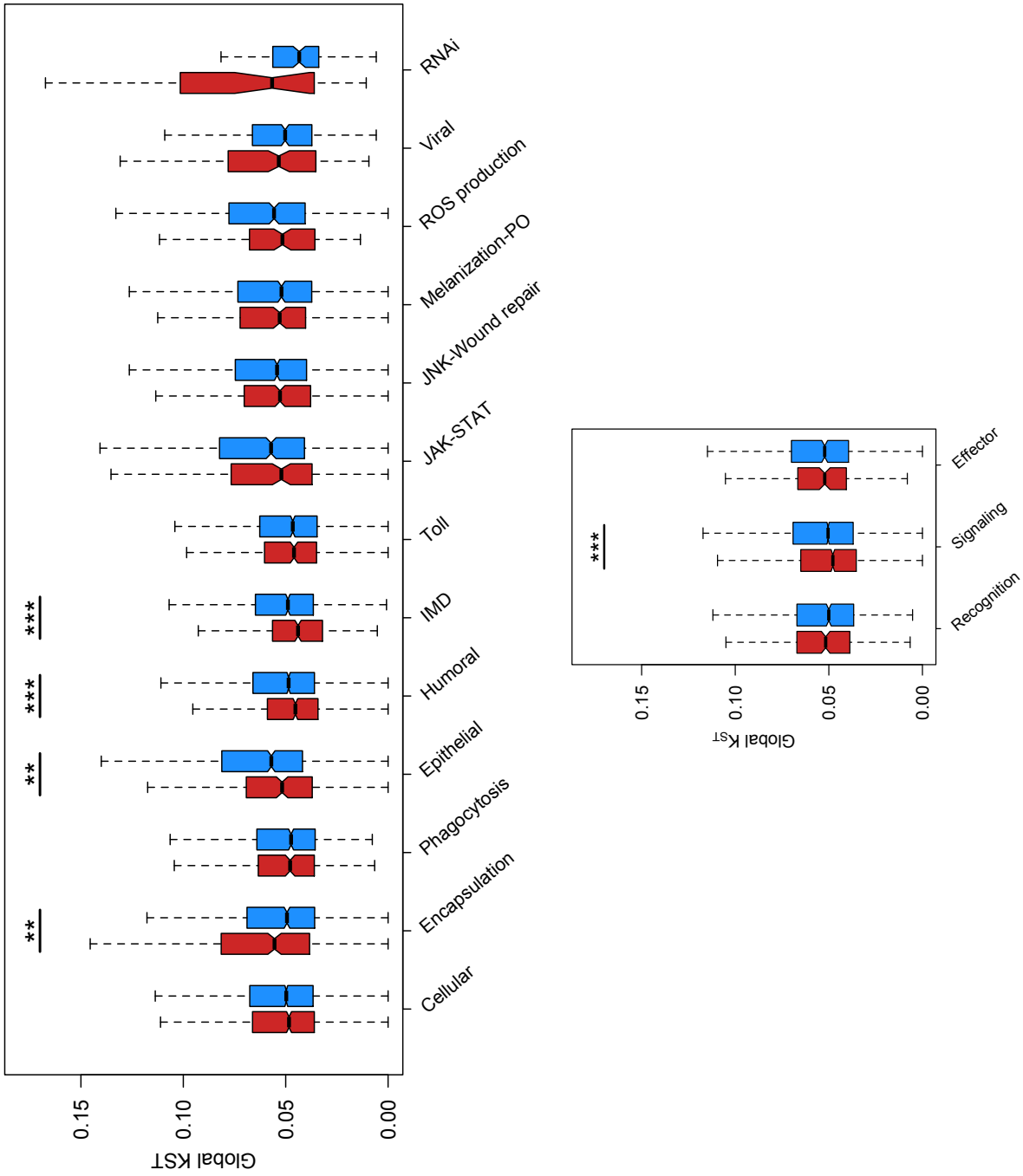
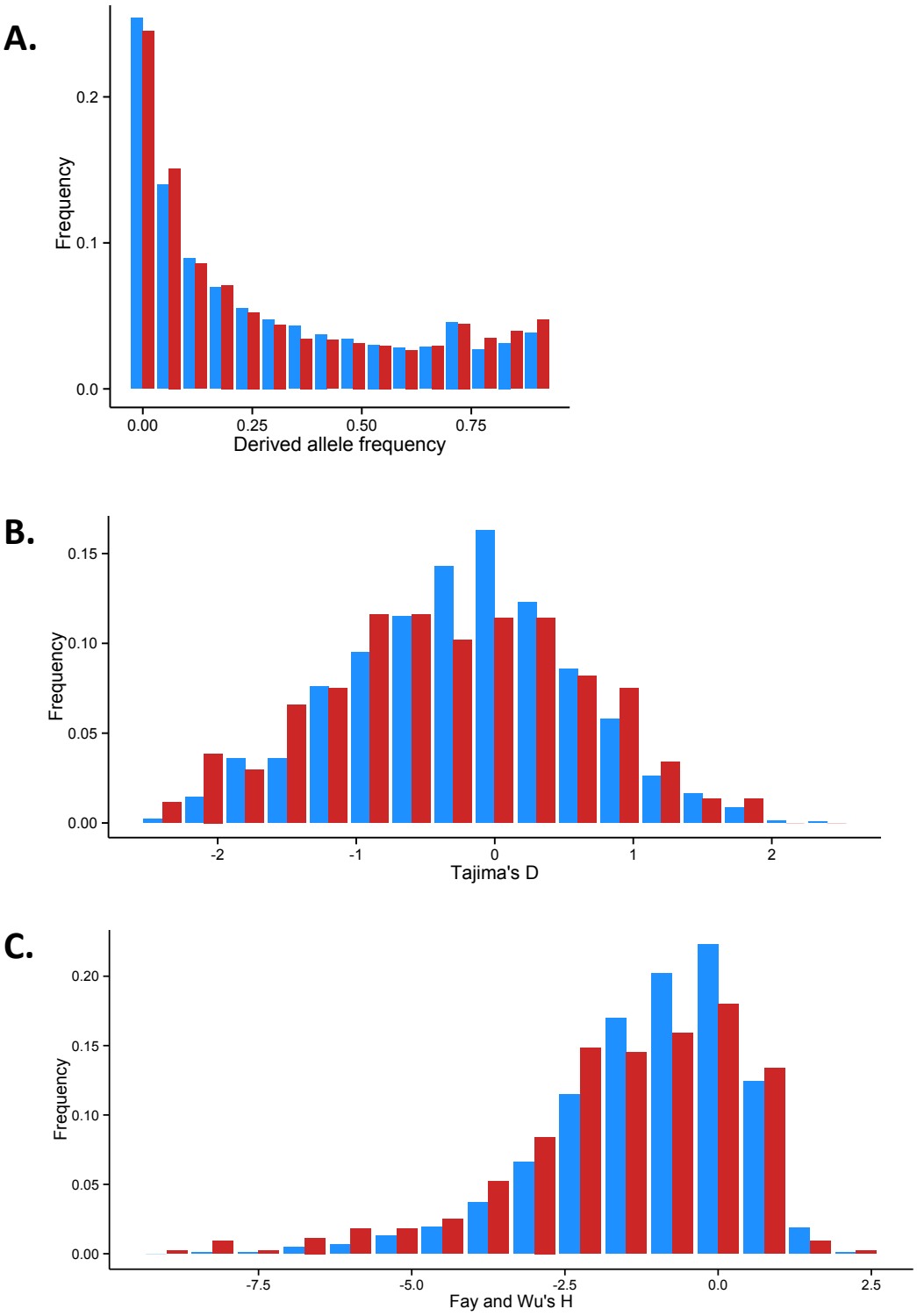

RNAi
Viral
ROS production
Melanization-PO
JNK-Wound repair
JAK-STAT
Toll
IMD
Humoral
Epithelial
Phagocytosis
Encapsulation
Cellular
**
**
***
***
0.15
0.10
0.05
0.00
Global KST
Effector
Signaling
Recognition
***
0.15
0.10
0.05
0.00
Global KST

Figure 3.5. Evidence for selection on genes involved in encapsulation.
(A) In Tasmania, encapsulation genes (red) contain a higher proportion of high frequency derived SNPs relative to control genes (blue; K-S test, $P < 0.05$). Encapsulation genes also show more extreme negative values of (B) Tajima's D and (C) Fay and Wu's H as well as more overall variance in these distributions (Tajima's D: Mann-Whitney U test, P =N.S.; F test $P < 0.01$; Fay and Wu's H: Mann-Whitney U test, $P < 0.01$.; Levene's test, $P < 0.05$. Significance values are Bonferroni corrected for multiple testing.) Tajima's D and Fay and Wu's H were calculated within 1-kb windows. Pictured here are values from Tasmania, but Beijing, Netherlands, and New York populations showed similar trends (Appendix 3).


A.
Frequency
0.2
0.1
0.0
0.00
0.25
0.50
0.75
Derived allele frequency
B.
Frequency
0.15
0.10
0.05
0.00
-2
-1
0
1
2
Tajima's D
C.
Frequency
0.20
0.15
0.10
0.05
0.00
-7.5
-5.0
-2.5
0.0
2.5
Fay and Wu's H

cant for four out of the eight distributions (Figure 3.5; Tajima's *D* in Beijing and the Netherlands; Fay and Wu's *H* in Ithaca and Tasmania; $P < 0.05$). We also observed a trend where these statistics not only had a lower mean, but also showed greater variance in encapsulation genes relative to controls, suggesting that a smaller proportion of windows in encapsulation genes are evolving neutrally (Figure 3.5; Tajima's *D*: the Netherlands and Tasmania, F test, $P < 0.01$; Fay and Wu's *H*: Beijing, Ithaca, and Tasmania, Levene's test, $P < 0.05$.) Like with F_{ST} and K_{ST}, these two statistics showed similar, but less significant patterns in the melanization-PO genes. Only Fay and Wu's *H* in the Netherlands showed a significant leftward shift after correction for multiple testing ($P < 0.001$).

Viral defense genes did not deviate from controls in this initial analysis. However, when we limited the analysis to windows that included at least 500 bp of coding sequence, we detected a pattern of high population differentiation (global K_{ST}) in viral defense genes (Mann-Whitney U test, $P < 0.05$).

Patterns in additional immune pathways

While encapsulation, melanization, and viral defense genes, showed global patterns of populations differentiation, these were not the only instances of deviations from control genes. The JNK/wound-repair genes showed increased population structure in comparisons between Beijing and the other three derived populations ($P < 0.005$). In several instances, the unfolded SFS of a pathway was skewed towards more high frequency derived alleles (K-S Test, $P < 0.05$ after Bonferroni correction). In the Netherlands, Ithaca, and Tasmania, cellular response genes were enriched for high-frequency derived alleles. Conversely, the SFS of cellular response genes in Beijing was left-skewed. In the Netherlands and Ithaca, humoral response

genes and JNK/wound-repair genes had a right-skewed SFS. This represents one of the few instances where we detected evidence for local selection on humoral genes. Overall, the Netherlands had the greatest number of pathways (six) that significantly deviated from controls. While the ecological reasons for this are unclear, the pattern suggests that the Netherlands flies may have experienced more extreme selection than the other populations. Conversely, no immune pathway had a significantly skewed SFS in the Zimbabwe population. This suggests that the patterns of selection in the derived populations arose from flies encountering novel environments, not from a universal process of host-pathogen co-evolution. In the latter instance, we would have likely seen evidence for heightened adaptation in Zimbabwe as well.

Signaling genes show reduced levels of population differentiation

We also observed differences among functional classes (*e.g.* recognition, signaling, and effector). Of the three major classes, signaling genes were the least differentiated among populations. Within 1-kb windows no pairwise F_{ST} or global K_{ST} comparison suggested heightened population differentiation and global K_{ST} was significantly reduced in these genes compared to controls (Figure 3.4; $P < 0.0001$). Conversely, effector and recognition genes showed elevated F_{ST} in at least one population pair. Effector genes—in particular AMPs and ROS—showed high differentiation between Ithaca and Tasmania, with Ithaca also having a high S_{nn} for effectors. F_{ST} patterns across lysozyme genes trended high in all comparisons with Beijing, although only one (Beijing-Netherlands) remained significant after Bonferroni correction for multiple testing. Recognition genes displayed a high F_{ST} between the Netherlands and Tasmania. These patterns suggest that local immune adaptation has largely occurred through

selection at the ends of pathways—particularly at initial recognition genes—and not through changes in signaling genes.

Discussion

Demographic factors can leave strong genomic signatures, making inferences of selection difficult. Indeed, genome-wide analysis of these 84 fly lines has shown that the populations are genetically distinct (Arguello *et al.*, in prep), so determining which differentiated regions are due to drift and which are due to selection remains a difficult problem. Here, we have detected signatures of local adaptation in a discrete set of genes through the careful selection of local genomic controls. We performed inferences of selection by pairing each immune gene with four control genes matched for size, genome-position, and local recombination rate. This method allowed us to confidently identify genes and pathways that are evolving in ways that significantly differ from the background average. As it is a conservative approach, we do not detect all instances of selection. Rather, the patterns we describe are the strongest, not the only, occurrences of local adaptation in immune genes.

At the single gene level, few patterns emerged across populations. This suggests that local immune adaptation has occurred along slightly different routes in each population, although such a pattern could also be compounded by sampling stochasticity. On the pathway level, however, we found consistent signals suggesting broad patterns in pathogen distribution and diversity.

One notable observation involves genes that participate in parasitoid wasp defense through the processes of encapsulation, phenoloxidase production, and melanization. Genes in these pathways showed unusually high levels of population differentiation as well as other

signs of selection (low Tajima's *D* and Fay and Wu's *H*) in derived populations (Figure 3.5). These processes protect the fly from parasitoid wasp attack, providing evidence that parasitoid wasps—more so than other known fly pathogens—have shaped the recent evolutionary trajectory of the fly's immune genes.

Numerous parasitoid wasp species prey on *Drosophila* and can infect up to 80% of the flies in a population (Carton Y, et al. 1986). Some are known to use virulence factors that interfere with the fly's encapsulation response, leading to a potential 'Red Queen' scenario between the host and parasite (Labrosse, et al. 2003; Mortimer, et al. 2013; Mortimer, et al. 2012; Rizki and Rizki 1990). Phenotypic differences in *D. melanogaster* encapsulation rates have been observed globally as well as on a much finer geographic scale (Dupas, et al. 2009; Kraaijeveld and van Alphen 1995), and our genetic results provide an important counterpart to this known phenotypic structuring. In addition, the genetic basis for extant variation in encapsulation ability is just beginning to be dissected, and our analysis provides further candidate haplotypes that may contribute to population-level differences among flies (Table 3.2).

A second pathway displaying unusually high population structure was viral defense. In our test for polygenic selection, this pattern was only apparent in coding-enriched regions, but viral defense genes were also enriched in the top 1% and 5% of global K_{ST} values across all windows. Viral response genes, and in particular genes involved in RNAi, are among the fastest evolving genes across the *Drosophila* phylogeny (Kolaczkowski, et al. 2011; Obbard, et al. 2006; Obbard, et al. 2009). Functional studies of *D. melanogaster*-virus interactions have highlighted several instances of single alleles aiding resistance against viruses, and in some cases this resistance is virus-specific (Magwire, et al. 2011; Magwire, et al. 2012; Wilfert and Jiggins 2010). It is therefore not surprising that we also found strong signatures of popula-

tion differentiation within several individual viral defense genes (*CHKov1*/*CHKov2*, *egh*, and *armi*; Table 3.2).

As opposed to parasitoids and viruses, no bacterial or fungal pathogens are known to be specialists of *Drosophila*. While it is logistically difficult to collect infected flies in the wild, surveys have been conducted, and they have found only generalist pathogens (Juneja 2011). If *D. melanogaster* is in truth assailed by an array of generalist bacteria and fungi, we would expect little to no strong directional selection on the immune pathways that deal most directly with these assaults. Indeed, humoral response genes, and in particular the IMD pathway showed no signs of selection within individual populations. In addition, these genes displayed significantly reduced population structure globally (Figure 3.4) and in several population pair comparisons, suggesting that flies have not adapted to novel bacterial and fungal pathogens when moving to new environments. The blanket assumption that humoral genes are universally conserved across populations, however, is contradicted by the potential signs of selection we see on certain antimicrobial compounds (discussed below).

Past work has found evolutionary differences among genes based on their functional classification as recognition receptors, signaling molecules, or effector proteins (Sackton, et al. 2007). Our analyses of population differentiation suggest that both recognition and effector proteins play a role in local adaptation. Through our analysis of individual genes, we identified eight genes with high population differentiation that encode pathogen recognition proteins. All of these proteins function in the cellular response. In addition, considering all phagocytosis recognition receptors as a group, we detect heightened population differentiation in several population pairs as well as a marginally significant increase in global population structure (Mann-Whitney U test, $P = 0.01650$ before multiple testing correction). In their

analysis of immune gene evolution ascertained through tests of sequence divergence among species within the *melanogaster* group, Sackton et al. (2007) described a similar pattern, identifying ten receptor genes that showed signs of positive selection, nine of which putatively function in phagocytosis. With the exception of *pes* (a gene that displayed a nominal signature of positive selection in Sackton et al. (2007)), the genes identified by our two studies differ. Still, the similarity in our large-scale observations is striking and suggests that recognition proteins—particularly those that function in the cellular immune response—are subject to both temporally and spatially variable selection. The differences in the actual genes we identify may be due to slight differences in power or temporal and spatial fluctuations in which receptors experience selection at a certain time and location.

In contrast to what they saw for recognition genes, Sackton et al. (2007) found little evidence of positive selection acting in effector genes at the nucleotide level. In line with this observation, both our estimates of α and those of Obbard et al. (2009) were low for effector genes, and even negative for AMPs (-1.697 [-7.363, 1.412]). Effector genes contained unusually high nucleotide diversity, but this variation was not population-specific; only one population pair (Ithaca-Tasmania) showed elevated population differentiation. These observations seem to support the hypothesis that effector genes experience relaxed selection, possibly because of their redundancy; however, we do find individual instances where effectors may have undergone population-specific selection. Several effector genes (CG6426, CG8492, proPO59, and yellow-f) showed significantly elevated pairwise F_{ST}. In addition, although our test lacked the power to detect selection in *D. melanogaster*'s short AMP genes, the 1-kb window including the coding region for the AMP *Dpt* was in the top 1% of F_{ST} values in Beijing-Zimbabwe comparisons. While two major haplotypes of *Dpt* are segregating globally, we found only one

in Beijing, suggesting that strong selection favored this single allele in this one population. Recently, these two haplotypes have been associated with variation in resistance to certain bacterial pathogens (Unckless et al in prep). While it is unknown whether *P. rettgeri* infects flies in east Asia, this laboratory experiment demonstrates a phenotypic difference between these alleles and gives biological support for the evolutionary signals we see.

In addition, other genetic processes may contribute to effector gene adaptation. While there is little evidence for long-term adaptation at the nucleotide level, effector gene families expand and contract at an unusually rapid pace (Sackton et al. 2007). Our focus on single nucleotide variants does not reflect this source of genetic novelty, but we note that these fly lines do contain two segregating duplications in the drosomycin family, a group of anti-fungal AMPs (M. Cardoso-Moreira, pers. comm.). The adaptive role played by such copy number variants remains to be determined at the population level.

In conclusion, it is important to note that we have explored only a subset of the genes that affect *D. melanogaster* fitness. We chose to focus on genes that are most directly involved with the immune response. Expression studies, however, have shown that a much wider set of genes is activated during experimental infection (De Gregorio, et al. 2001; Wertheim, et al. 2005), and additional genes have even been identified during artificial selection experiments (Wertheim, et al. 2011). While some of these changes reflect a simple perturbation of homeostasis, others no doubt play a role in aiding the fly's ability to survive infection. With this in mind, it is possible that these flies have adapted to local bacterial or fungal pathogens through non-canonical responses that were not included in our study (Kacsoh, et al. 2013). In addition, while treated here as distinct, these pathways communicate with one another, and in many instances, there are multiple mechanisms through which any single pathway can be triggered.

Continuing functional studies along with expanded geographic sampling of natural pathogens will allow us to more fully assess the process of adaptation by *Drosophila* to pathogenic infection.

Acknowledgements

Roman Arguello, Jen Grenier, Margarida Cardoso Moreira, and Srikanth Gottipati performed the sequencing and variant calling in these lines.

References

Ayres JS, Freitag N, Schneider DS 2008. Identification of Drosophila mutants altering defense of and endurance to Listeria monocytogenes infection. Genetics 178: 1807-1815. doi: 10.1534/genetics.107.083782

Ayres JS, Schneider DS 2009. The role of anorexia in resistance and tolerance to infections in Drosophila. Plos Biology 7: e1000150. doi: 10.1371/journal.pbio.1000150

Babin A, Kolly S, Schneider F, Dolivo V, Zini M, Kawecki TJ 2014. Fruit flies learn to avoid odours associated with virulent infection. Biol Lett 10: 20140048. doi: 10.1098/rsbl.2014.0048

Carton Y, Bouletreau M, van Alphen JJ, van Lenteren JC. 1986. The Drosophila parasitic wasps. In: Ashburner M, Carson L, Thompson JN, editors. The genetics and biology of Drosophila. London: Academic Press. p. 347-394.

Chavez-Galarza J, Henriques D, Johnston JS, Azevedo JC, Patton JC, Munoz I, De la Rua P, Pinto MA 2013. Signatures of selection in the Iberian honey bee (Apis mellifera iberiensis) revealed by a genome scan analysis of single nucleotide polymorphisms. Molecular Ecology 22: 5890-5907. doi: 10.1111/mec.12537

Comeron JM, Ratnappan R, Bailin S 2012. The many landscapes of recombination in Drosophila melanogaster. Plos Genetics 8: e1002905. doi: 10.1371/journal.pgen.1002905

Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, Sagnon N, Lazzaro BP 2010. De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. Plos One 5: e14202. doi: 10.1371/journal.pone.0014202

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L 2013. Evidence for polygenic adaptation to pathogens in the human genome. Molecular Biology and Evolution 30: 1544-1558. doi: 10.1093/molbev/mst080

David JR, Capy P 1988. Genetic-Variation of Drosophila-Melanogaster Natural-Populations. Trends in Genetics 4: 106-111. doi: Doi 10.1016/0168-9525(88)90098-4

De Gregorio E, Spellman PT, Rubin GM, Lemaitre B 2001. Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc Natl Acad Sci U S A 98: 12590-12595. doi: 10.1073/pnas.221458698

Dupas S, Dubuffet A, Carton Y, Poirie M 2009. Local, geographic and phylogenetic scales of coevolution in Drosophila-parasitoid interactions. Adv Parasitol 70: 281-295. doi: 10.1016/S0065-308X(09)70011-9

Erler S, Lhomme P, Rasmont P, Lattorff HM 2014. Rapid evolution of antimicrobial peptide genes in an insect host-social parasite system. Infect Genet Evol 23: 129-137. doi: 10.1016/j.meegid.2014.02.002

Eyre-Walker A, Keightley PD 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution 26: 2097-2108. doi: 10.1093/molbev/msp119

Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlotterer C, Flatt T 2012. Genome-wide patterns of latitudinal differentiation among populations of Drosophila melanogaster from North America. Molecular Ecology 21: 4748-4769. doi: 10.1111/j.1365-294X.2012.05731.x

Ferrandon D, Imler JL, Hetru C, Hoffmann JA 2007. The Drosophila systemic immune

response: sensing and signalling during bacterial and fungal infections. Nat Rev Immunol 7: 862-874. doi: 10.1038/nri2194

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA 2010. Drosophila melanogaster recombination rate calculator. Gene 463: 18-20. doi: Doi 10.1016/J.Gene.2010.04.015

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. Plos Genetics 7: e1002355. doi: 10.1371/journal.pgen.1002355

Gibson G 2012. Rare and common variants: twenty arguments. Nature Reviews Genetics 13: 135-145. doi: 10.1038/nrg3118

Howick VM, Lazzaro BP 2014. Genotype and diet shape resistance and tolerance across distinct phases of bacterial infection. BMC Evol Biol 14: 56. doi: 10.1186/1471-2148-14-56

Hubner S, Rashkovetsky E, Kim YB, Oh JH, Michalak K, Weiner D, Korol AB, Nevo E, Michalak P 2013. Genome differentiation of Drosophila melanogaster from a microclimate contrast in Evolution Canyon, Israel. Proc Natl Acad Sci U S A 110: 21059-21064. doi: 10.1073/pnas.1321533111

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C, Broad Institute Genome Sequencing P, Whole Genome Assembly T, Baldwin J, Bloom T, Jaffe DB, Nicol R, Wilkinson J, Lander ES, Di Palma F,

Lindblad-Toh K, Kingsley DM 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55-61. doi: 10.1038/nature10944

Juneja P 2011. Short Term Evolution In The Immune Response Of Drosophila Melanogaster: Insights From Studies Of Population Genetics And The Epidemiology Of Bacterial Infection [dissertation]. Cornell University.

Juneja P, Lazzaro BP 2010. Haplotype structure and expression divergence at the Drosophila cellular immune gene eater. Molecular Biology and Evolution 27: 2284-2299. doi: 10.1093/molbev/msq114

Kacsoh BZ, Lynch ZR, Mortimer NT, Schlenke TA 2013. Fruit flies medicate offspring after seeing parasites. Science 339: 947-950. doi: 10.1126/science.1229625

Keller A 2007. Drosophila melanogaster's history as a human commensal. Current Biology 17: R77-R81. doi: Doi 10.1016/J.Cub.2006.12.031

Kolaczkowski B, Hupalo DN, Kern AD 2011. Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. Molecular Biology and Evolution 28: 1033-1042. doi: 10.1093/molbev/msq284

Kraaijeveld AR, Layen SJ, Futerman PH, Godfray HC 2012. Lack of phenotypic and evolutionary cross-resistance against parasitoids and pathogens in Drosophila melanogaster. Plos One 7: e53002. doi: 10.1371/journal.pone.0053002

Kraaijeveld AR, Limentani EC, Godfray HC 2001. Basis of the trade-off between parasitoid resistance and larval competitive ability in Drosophila melanogaster. Proc Biol Sci 268: 259-261. doi: 10.1098/rspb.2000.1354

Kraaijeveld AR, van Alphen JJ 1995. Geographical variation in encapsulation ability of *Drosophila melanogaster* larvae and evidence for parasitoid-specific components. Evolutionary Ecology 9: 10-17.

Labrosse C, Carton Y, Dubuffet A, Drezen JM, Poirie M 2003. Active suppression of D. melanogaster immune response by long gland products of the parasitic wasp Leptopilina boulardi. J Insect Physiol 49: 513-522.

Lamichhaney S, Martinez Barrio A, Rafati N, Sundstrom G, Rubin CJ, Gilbert ER, Berglund J, Wetterbom A, Laikre L, Webster MT, Grabherr M, Ryman N, Andersson L 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proc Natl Acad Sci U S A 109: 19345-19350. doi: 10.1073/pnas.1216128109

Laurent SJ, Werzner A, Excoffier L, Stephan W 2011. Approximate Bayesian analysis of Drosophila melanogaster polymorphism data reveals a recent colonization of Southeast Asia. Molecular Biology and Evolution 28: 2041-2051. doi: 10.1093/molbev/msr031

Lazzaro BP, Flores HA, Lorigan JG, Yourth CP 2008. Genotype-by-environment interactions and adaptation to local temperature affect immunity and fecundity in Drosophila melanogaster. Plos Pathogens 4: e1000025. doi: 10.1371/journal.ppat.1000025

Lazzaro BP, Little TJ 2009. Immunity in a variable world. Philos Trans R Soc Lond B Biol Sci 364: 15-26. doi: 10.1098/rstb.2008.0141

Lazzaro BP, Sackton TB, Clark AG 2006. Genetic variation in Drosophila melanogaster resistance to infection: a comparison across bacteria. Genetics 174: 1539-1554. doi: 10.1534/genetics.105.054593

Loytynoja A, Goldman N 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102: 10557-10562. doi: 10.1073/pnas.0409137102

Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM 2011. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. Plos Genetics 7: e1002337. doi: 10.1371/journal.pgen.1002337

Magwire MM, Fabian DK, Schweyen H, Cao C, Longdon B, Bayer F, Jiggins FM 2012. Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in Drosophila melanogaster. Plos Genetics 8: e1003057. doi: 10.1371/journal.pgen.1003057

McKean KA, Yourth CP, Lazzaro BP, Clark AG 2008. The evolutionary costs of immunological maintenance and deployment. BMC Evol Biol 8: 76. doi: 10.1186/1471-2148-8-76

McTaggart SJ, Obbard DJ, Conlon C, Little TJ 2012. Immune genes undergo more adaptive evolution than non-immune system genes in Daphnia pulex. BMC Evol Biol 12: 63. doi: 10.1186/1471-2148-12-63

Mortimer NT, Goecks J, Kacsoh BZ, Mobley JA, Bowersock GJ, Taylor J, Schlenke TA 2013. Parasitoid wasp venom SERCA regulates Drosophila calcium levels and inhibits cellular immunity. Proc Natl Acad Sci U S A 110: 9427-9432. doi: 10.1073/pnas.1222351110

Mortimer NT, Kacsoh BZ, Keebaugh ES, Schlenke TA 2012. Mgat1-dependent N-glycosylation of membrane components primes Drosophila melanogaster blood

cells for the cellular encapsulation response. Plos Pathogens 8: e1002819. doi: 10.1371/journal.ppat.1002819

Obbard DJ, Jiggins FM, Halligan DL, Little TJ 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. Current Biology 16: 580-585. doi: 10.1016/j.cub.2006.01.065

Obbard DJ, Welch JJ, Kim KW, Jiggins FM 2009. Quantifying adaptive evolution in the Drosophila immune system. Plos Genetics 5: e1000698. doi: 10.1371/journal.pgen.1000698

Pespeni MH, Garfield DA, Manier MK, Palumbi SR 2012. Genome-wide polymorphisms show unexpected targets of natural selection. Proc Biol Sci 279: 1412-1420. doi: 10.1098/rspb.2011.1823

Quintana-Murci L, Clark AG 2013. Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol 13: 280-293. doi: 10.1038/nri3421

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rizki RM, Rizki TM 1990. Parasitoid virus-like particles destroy Drosophila cellular immunity. Proc Natl Acad Sci 87: 8388-8392.

Rockman MV 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution 66: 1-17. doi: 10.1111/j.1558-5646.2011.01486.x

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG 2007. Dynamic evolution of the innate immune system in Drosophila. Nature Genetics 39: 1461-1468. doi: 10.1038/ng.2007.60

Salazar-Jaramillo L, Paspati A, van de Zande L, Vermeulen CJ, Schwander T, Wertheim B 2014. Evolution of a cellular immune response in Drosophila: a phenotypic and genomic comparative analysis. Genome Biology and Evolution 6: 273-289. doi: 10.1093/gbe/evu012

Savolainen O, Lascoux M, Merila J 2013. Ecological genomics of local adaptation. Nature Reviews Genetics 14: 807-820. doi: 10.1038/nrg3522

Scheinfeldt LB, Tishkoff SA 2013. Recent human adaptation: genomic approaches, interpretation and insights. Nature Reviews Genetics 14: 692-702. doi: 10.1038/nrg3604

Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J 2010. Adaptation genomics: the next generation. Trends Ecol Evol 25: 705-712. doi: 10.1016/j.tree.2010.09.002

Thornton K, Andolfatto P 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172: 1607-1619. doi: 10.1534/genetics.105.048223

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV 2010. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. Nature Genetics 42: 260-263. doi: 10.1038/ng.515

Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, Dong Y, Jiang H, Kanost MR, Koutsos AC, Levashina EA, Li J, Ligoxygakis P, Maccallum RM, Mayhew GF, Mendes A, Michel K, Osta MA, Paskewitz S, Shin SW, Vlachou D, Wang L, Wei W, Zheng L, Zou Z, Severson DW, Raikhel AS, Kafatos FC, Dimopoulos G, Zdobnov EM,

Christophides GK 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316: 1738-1743. doi: 10.1126/science.1139862

Weir BS, Cockerham CC 1984. Estimating F-Statistics for the Analaysis of Population Structure. Evolution 38: 1358-1370.

Wertheim B, Kraaijeveld AR, Hopkins MG, Walther Boer M, Godfray HC 2011. Functional genomics of the evolution of increased resistance to parasitism in Drosophila. Molecular Ecology 20: 932-949. doi: 10.1111/j.1365-294X.2010.04911.x

Wertheim B, Kraaijeveld AR, Schuster E, Blanc E, Hopkins M, Pletcher SD, Strand MR, Partridge L, Godfray HC 2005. Genome-wide gene expression in response to parasitoid attack in Drosophila. Genome Biol 6: R94. doi: 10.1186/gb-2005-6-11-r94

Wilfert L, Jiggins FM 2010. Disease association mapping in Drosophila can be replicated in the wild. Biol Lett 6: 666-668. doi: 10.1098/rsbl.2010.0329

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871. doi: 10.1093/bioinformatics/btp394

CHAPTER 4

Network structure constrains the genetic architecture of local adaptation in *Drosophila melanogaster*

Abstract

The interactions among genes within regulatory networks are known to affect molecular patterns of species divergence, yet their effects on local adaptation are not well understood. Here, we explore how network-centrality, number of interacting partners, and tissue expression breadth correlate with measures of genetic differentiation and adaptation within immune genes across five populations of *Drosophila melanogaster*. Our study supports a model in which genetic interactions amplify the strength and efficacy of purifying selection in functionally important regions, and thereby constrain the evolutionary divergence of genes embedded within interaction networks. Signatures of constraint are most apparent in parasitoid defense genes, which were previously shown to bear signs of recent positive selection. For genes experiencing weaker selection, constraint appears to be localized and may not come at the expense of adaptation occurring in other regions of the gene. Together, these results demonstrate that molecular genetic proxies for pleiotropy successfully predict general patterns of evolutionary constraint, and that their utility is amplified when combined with measures of function and selection.

Introduction

Proteins function and evolve not in isolation, but in constant interplay with extra- and intra-organismal components. In many cases, the outcomes of these interactions are readily apparent, but when genes have multiple pleiotropic effects, the evolutionary consequences are less clear

(Hill and Zhang 2012a, b; Wagner and Zhang 2011; Wang, et al. 2010). For instance, theory predicts that pleiotropy will constrain evolutionary rates in certain instances, but not universally limit adaptation (Wagner, et al. 2008) or the evolution of complexity (Wang, et al. 2010) in the organism as a whole.

Empirical evidence of pleiotropic effects lags behind our theoretical understanding. Largely this is because of difficulties with assessing the full phenotypic effects—and therefore pleiotropy—of a gene. On the molecular level, several commonly used network-level statistics are thought to correlate with extent of pleiotropy. These measures, which include network-centrality, tissue expression bias, and degree of interaction, can be used as proxies for a gene's pleiotropic effects. On long evolutionary time-scales there are well-documented correlations between these metrics and a protein's rate of evolution. Fraser, et al. (2002) first noted that there is a negative correlation between a protein's rate of adaptation and its number of interacting partners. As studies were expanded to include increased knowledge about network structure, it was found that a protein's centrality largely constrains its evolution (Fraser 2005; Hahn and Kern 2005). Most recently, a comparative genetics analysis across *Drosophila* species showed that tissue-biased expression also predicts evolutionary rates (Larracuente, et al. 2008). Genes with more limited expression tend to have a more limited function. This frees them from constraint and allows them to adapt more rapidly.

In humans, these observations have been extended to the population level (Casals, et al. 2011; Luisi, et al. 2012). In other organisms, however, little work has been done to determine whether these same genetic factors constrain routes of short-term adaptation (but see Alvarez-Ponce, et al. 2012; Vishnoi, et al. 2011). In addition, the larger context of these factors has rarely been explored, and dichotomies are sometimes established instead of studying

the interplay between directional selection and constraint (Opulente, et al. 2013). We are now at a point where we can combine information about the genetic architecture underlying complex traits with full genome data sets describing extant population diversity. This combined knowledge will allow us to broaden our understanding of how local adaptation progresses at the genetic level.

Here, we test whether pleiotropy plays a role in local adaptation by analyzing patterns of selection and constraint across five populations of *D. melanogaster*. Drawing from our previous in-depth analysis of immune genes (Chapter 3), we look for evidence that the position of a gene within its regulatory network places constraints on the routes of local immune adaptation. We find that constraint is most readily detected in genes undergoing strong directional selection. Genes that have more interacting partners experience greater purifying selection, but in the case of weaker directional selection, this constraint is not detected across the entire length of the gene.

Methods

Genomic data

Previously, we assembled a list of 375 immune genes with well-supported immune function (Chapter 3). We selected these genes through literature searches, seeking evidence of immune involvement beyond GO annotations. For each gene, we obtained information on nucleotide polymorphism from 84 sequenced inbred fly lines from five populations: Zimbabwe, the Netherlands, Beijing, Tasmania, and Ithaca, New York (The Global Diversity Lines panel; Grenier, et al. in prep). We masked genomic regions with poor "callability" (as defined in Grenier, et al., in prep) to ensure that our analyses excluded genomic regions with large amounts

of missing data. At sites that were heterozygous in a given line, we randomly sampled a single allele to use in all downstream analyses.

Population genetic analyses

As described in Chapter 3, we used custom Perl scripts to calculate measures of pairwise nucleotide diversity (π), selection (Tajima's *D* and Fay and Wu's *H*) and population divergence (K_{ST} and F_{ST}) for each gene. We also calculated these statistics within fixed 1-kb windows that covered the length of each gene. To calculate F_{ST}, we used the method of Weir and Cockerham (1984) and only made comparisons between a population and its most recent ancestral population (Grenier et al, in prep; Beijing-Zimbabwe, Netherlands-Zimbabwe, Ithaca-Netherlands, and Tasmania-Netherlands). We accounted for missing data by adjusting the sample size at each SNP and then calculated F_{ST} across regions by separately averaging the numerator and denominator (Weir and Cockerham 1984). Negative values were rounded to zero. For each gene, we calculated GC-content using custom Perl scripts, and obtained estimates of local recombination rate with the *Drosophila melanogaster* Recombination Rate Calculator v 2.3 (Fiston-Lavier, et al. 2010).

Divergence analyses

For 292 genes, we were able to identify one-to-one *D. simulans* orthologs using FlyBase (v. FB2013_06). We then downloaded all available transcripts for these genes from FlyBase and performed all possible *D. melanogaster-D. simulans* alignments using PRANK (Loytynoja and Goldman 2005). As described in Chapter 3, we selected the longest transcript pair with the fewest number of nonsynonymous substitutions to represent the gene. With custom Perl scripts, we used these alignments to calculate synonymous (*ds*) and nonsynonymous

divergence (*dn*) using the Nei-Gojobori method with a Jukes-Cantor correction (Nei and Gojobori 1986).

Network-level analyses

We downloaded all *D. melanogaster* interactions available on BioGrid (v. 3.2.110; Chatr-Aryamontri, et al. 2013). In total, the network contained 8,702 genes, 66,222 protein-protein interactions, and 23,074 genetic interactions. In our analyses, we considered the combined network of both genetic and protein-protein interactions as well as a more limited network of only protein-protein interactions (PPI). Both networks gave similar results; values reported in the text are for the PPI network. For each immune protein, we calculated its degree by counting the total number of interacting partners and classified these partners as immune or non-immune. Betweenness centrality of each protein in the interaction network was calculated using the betweenness_centrality function in the Python module NetworkX (Hagberg, et al. 2008). This measure is based on the number of times the shortest path connecting all pairs of proteins passes through the focal protein.

We obtained whole fly and tissue-specific expression data from FlyAtlas (Chintapalli, et al. 2007) and annotated the probes using Affymetrix Drosophila_2, Release 34 annotations. To determine the degree of tissue-biased expression, we calculated τ as:

$$\tau = \frac{\sum_{i=1}^{N} 1 - \frac{\log(S_i)}{\log(S_{\max})}}{N-1}$$

where S_i is the signal intensity in tissue i, S_{max} is the maximum intensity observed in any tissue, and *N* is the total number of tissues (Larracuente, et al. 2008). For adults, we used 9

somatic tissues: brain, crop, midgut, hindgut, Malpighian tubule, thoracico-abdominal ganglia, salivary gland, fat body, and eye. For larvae we used 7 tissues: tubule, fat body, salivary gland, midgut, hindgut, central nervous system, and trachea. Only probes detected in at least two replicate microarrays were counted as expressed in a given tissue. When multiple probes were present for a given gene, we averaged the average signal at each probe to calculate gene expression.

All statistical analyses were conducted in R (R Development Core Team 2011). Partial correlations were calculated with the package ppcor.

Results and Discussion

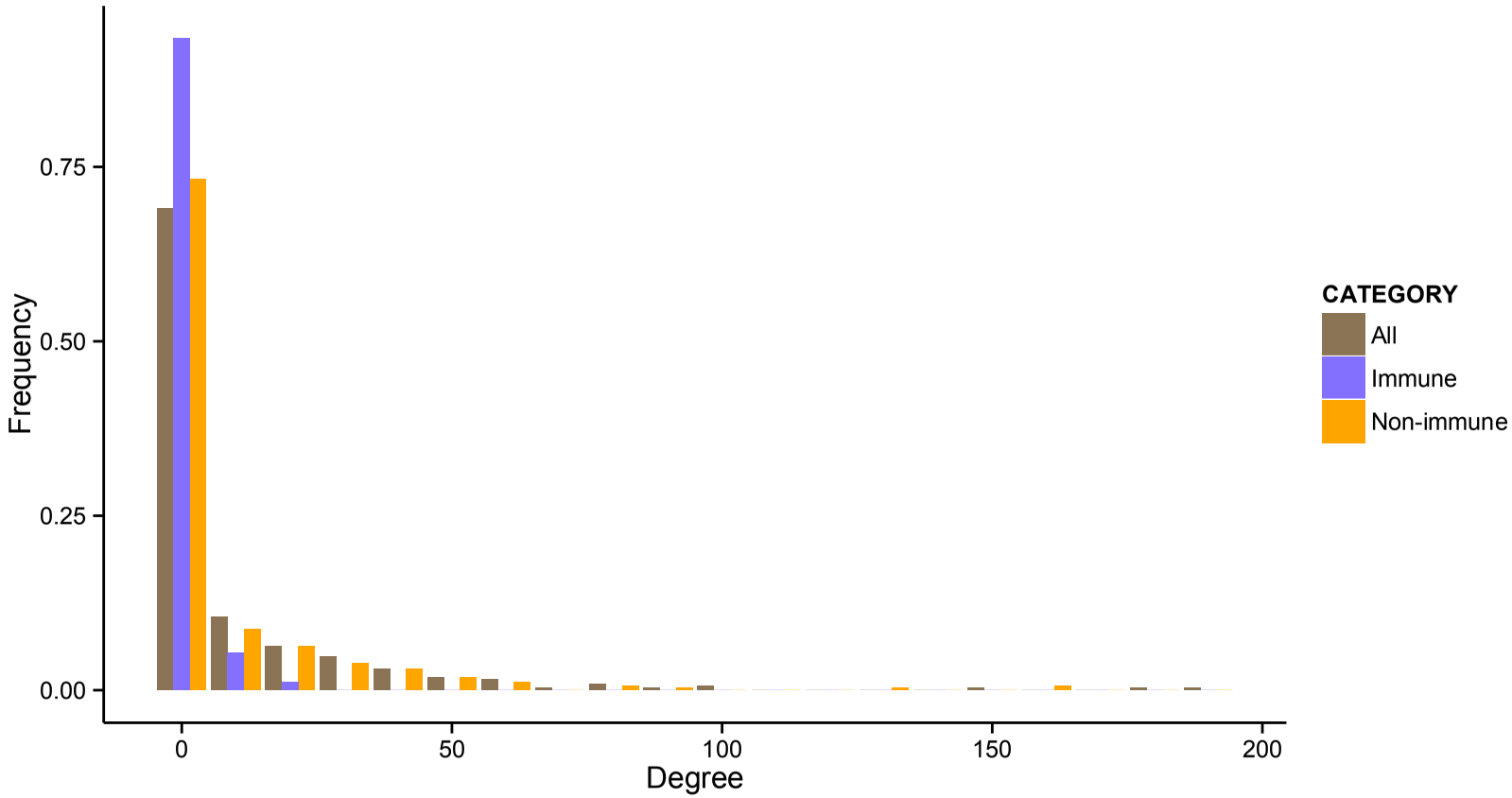
Proxies of genetic constraint vary across immune genes

For the purpose of our study, we assembled a hand-curated list of 375 *D. melanogaster* immune genes, drawn from the extensive *Drosophila* immune system literature. For each gene, we calculated three statistics that are proxies for genetic pleiotropy: degree, betweenness centrality and tissue-biased expression (τ). Degree was calculated from the number of known protein-protein interactions in the *D. melanogaster* gene network. Betweenness centrality is a measurement of a protein's placement within a network and describes the extent to which it acts as a hub. Tissue-biased expression can serve as a proxy for genetic pleiotropy as genes with more limited expression are likely to fulfill a more specialized biological function. For all three statistics, we found that immune genes displayed substantial heterogeneity (Figure 4.1).

Evidence for genetic constraint in a set of rapidly evolving immune genes

Previous studies of *Drosophila*'s response to parasitoid wasps have found that flies show significant variation in their level of defense (Dupas, et al. 2009). In a previous analysis of

Figure 4.1. Number of protein-protein interactions (degree) varies across immune genes. Immune degree measures the number of interactions occurring between two genes in our gene set. Non-immune degree measures interactions between an immune gene and a non-immune gene.

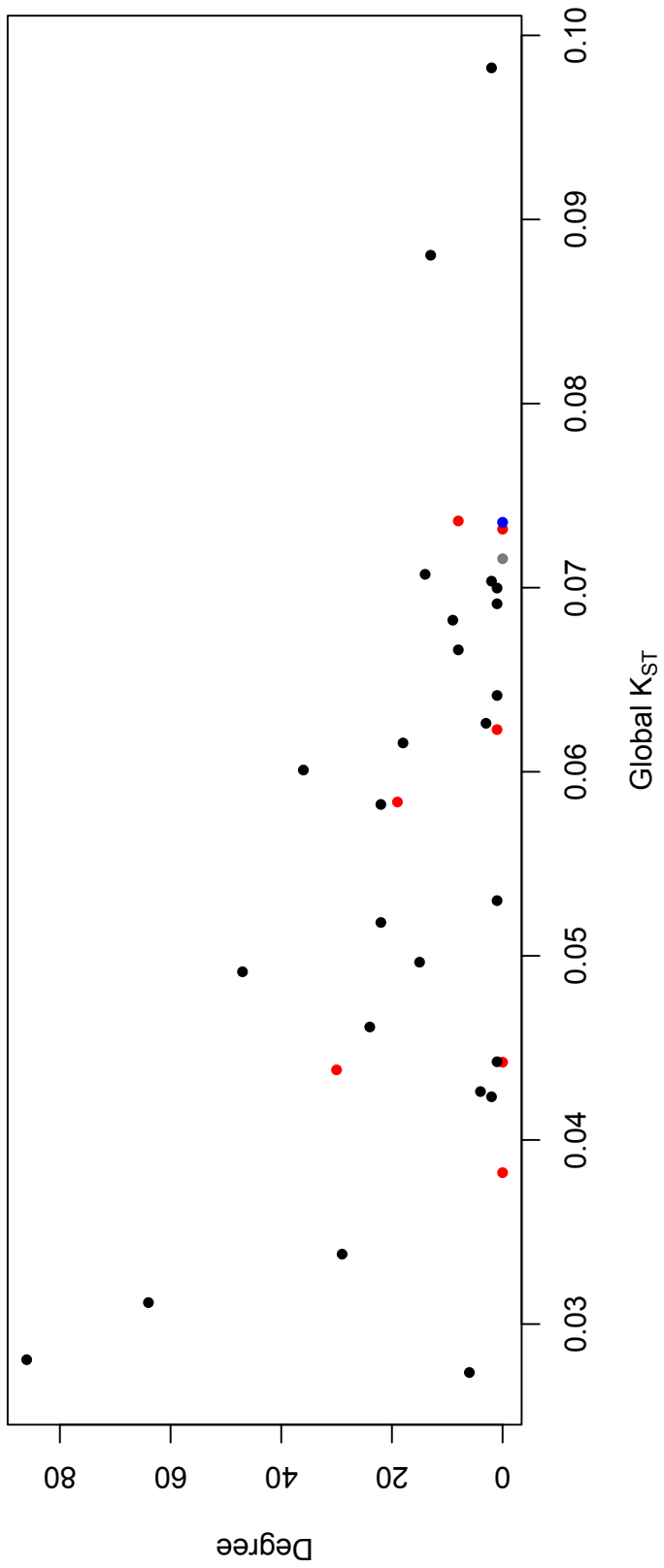

0.75
0.50
0.25
0.00
Frequency
0
50
100
150
200
Degree
CATEGORY
All
Immune
Non-immune

immune genes, we found that genes involved in defense against parasitoid wasps (namely, encapsulation genes) show a similar pattern. As a group, they display greater levels of population differentiation and stronger signals of selection relative to other immune genes (Chapter 3). Together, these results provide strong evidence that parasitoid defense is a trait that is experiencing significant selection pressure and undergoing local adaptation. We were therefore interested in determining the extent to which parasitoid defense genes displayed heterogeneity in their rate of local adaptation, and whether this heterogeneity correlated with measurements of pleiotropy or network structure.

We found a significant negative correlation between the degree of a protein and its population differentiation as measured by global K_{ST} (Figure 4.2; Spearman's $\rho = -0.4480$, $P = 0.01911$). Performing a partial correlation that accounted for gene length and GC-content at first and third codon positions, this relationship remained significant and strong (Spearman's partial $\rho = -0.4384$, $P = 0.02213$). No significant correlations were detected with betweenness centrality or τ.

It is possible that aspects of host-parasite co-evolution rather than network structure drove the correlation with degree. For instance, immune proteins that interact most directly with pathogens (namely, recognition and effector proteins) display faster rates of evolution (Sackton, et al. 2007). These same proteins tend to occupy peripheral positions on the immune network and have lower degree values. If these genes also displayed more rapid rates of population differentiation, the negative correlation between degree and global K_{ST} might only reflect differences among these genes classes. We therefore examined the function of each protein and classified it as signaling, effector, or recognition. As seen in Figure 4.2, the genes displaying high population differentiation do not preferentially have effector or recognition

Figure 4.2. Measurements of population differentiation and degree for genes involved in parasitoid wasp defense.
After accounting for gene length and GC-content at first and second codon sites, the correlation between global K_{ST} and degree is significant (Spearman's $\rho = -0.4308$, $P = 0.01015$). Points are colored according to protein function: black=signaling, red=recognition, blue=effector, grey=unknown


Degree
0
20
40
60
80
Global K_{ST}
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10

functions. In fact, when we limit our analysis to just signaling genes (*ie*, genes in the internal part of the network), the partial correlation between global K_{ST} and degree remains significant (Spearman's partial correlation controlling for gene length and GC-content at first and third codon positions, $\rho = -0.4535$, $P = 0.02286$).

Interacting proteins often show correlated rates of evolution which can be driven through either physical interactions or shared biological function (Clark, et al. 2012). We reasoned that the types of interacting partners, and more specifically the biological functions of those partners, might partially determine the extent of adaptive constraint placed on a protein. Many immune genes are known to also have non-immune functions, but the full extent of each gene's pleiotropic role is unknown. To this end, we categorized each interacting pair as Immune-Immune or Immune-Other and with these classifications calculated two additional metrics: immune degree and non-immune degree. We recognize that our list of immune genes is not exhaustive and likely excludes some genes with unrecognized immune function. Still, the *Drosophila* immune response is among the most well described processes in the fly, so our list likely includes the major players in the immune system. We therefore made the assumption that genes excluded from our list have at most peripheral immune involvement relative to the genes we include. Genes with higher immune degree would then be more strongly linked to immune processes. If anything, our incomplete knowledge of the system should make our analysis more conservative as it makes our non-immune group more immune-like.

There was no partial correlation between immune degree and global K_{ST} (Spearman's partial correlation controlling for gene length and GC-content at first and third codon positions, $\rho = -0.02076$, $P = 0.7586$). The partial correlation between non-immune degree and global K_{ST}, however, was strong (Spearman's partial $\rho = -0.4637$, $P = 0.01409$). We lack the

phenotypic knowledge to definitively say that a gene's functional breadth correlates with its number of non-immune partners, but this is a reasonable assumption. These patterns therefore suggest that protein-protein interactions impose the greatest constraints on short-term local adaptation when proteins are involved with multiple biological processes. However, it is likely that immune-immune interactions are under-represented in our network. Most studies of protein-protein interactions have been performed under stress-free conditions. Because of this, many interactions between immune genes are likely missing, biasing immune degree downwards.

Gene groups that experience weaker selection show less evidence of constraint

To determine whether these patterns are more generalizable, we expanded our analysis to our full immune gene set. As a whole, immune genes do not show heightened population differentiation, but individual genes do show signs of local adaptation (Chapter 3). Since lower selection pressures are present throughout the full set of immune genes, we expected to see a pattern that was similar to, but weaker than what we observed for encapsulation genes. This is because the observed effect of the constraint is likely to be proportional to the strength of directional selection. Indeed, this is what we observed. For immune genes as a whole, the partial correlation between degree and global K_{ST} was significant after accounting for gene length and GC-content ($\rho = -0.1356$, $P = 0.02916$). The correlation was even stronger between non-immune degree and global K_{ST} ($\rho = -0.1705$, $P = 0.005832$), but was not significant for immune degree ($\rho = -0.0381$, $P = 0.54$).

As we expected, this correlation was weaker than what we observed for parasitoid defense genes ($\rho = -0.4535$). Again we excluded signaling and effector proteins from the analy-

sis to determine whether pathogen-interacting proteins drive this pattern. Using only signaling proteins, we again found that the correlations were even stronger for both total degree (Spearman's partial ρ = -0.1805, $P = 0.02068$) and non-immune degree (Spearman's partial ρ = -0.2280, $P = 0.003159$). This implies that the biological functions of the genes are not driving this pattern.

Genetic constraints strengthen signatures of conservation

The correlation between degree and population differentiation was stronger for parasitoid defense genes than for all immune genes. The likely cause for this difference is variation in selection pressure; the overall strong selection pressure on parasitoid defense genes made the signature of constraint more apparent. We reasoned that genes under weak selection still experience constraint due to pleiotropy or network structure, but that the signatures of this constraint may be limited to certain functional regions of a gene (for instance, the regions coding the interaction domains involved in a protein-protein interaction). Statistics calculated across the entire length of a gene may not detect these localized regions of constraint. We therefore looked for more fine-scale signals by calculating Tajima's *D*, Fay and Wu's *H*, pairwise F_{ST} and global K_{ST} within 1-kb windows that covered the length of our genes. For each statistic we then calculated the z-score so we could compare values across populations.

For each gene, we selected the windows with the strongest evidence of constraint by identifying the lowest F_{ST} and K_{ST} z-scores within the gene. Across all these lines, genic regions display a lower F_{ST} than putatively neutral regions, suggesting that they experience a certain degree of constraint that counteracts the effects of drift (Arguello et al, in prep.). Therefore, the extreme windows we chose displayed more homogeneity among populations

than would be expected under a neutral scenario. Using these windows, we performed partial correlations with degree, tissue-bias, and centrality that controlled for gene length and GC-content.

The patterns that emerged were similar to what we observed in our analysis of full gene data. Both minimum K_{ST} and minimum F_{ST} were significantly correlated with measures of genetic constraint, although the exact patterns differed. The correlation between minimum global K_{ST} and total degree did not reach a significance threshold of $\alpha = 0.05$ (Spearman's partial $\rho = -0.1285$, $P = 0.05522$), but the correlation with non-immune degree was significant (Spearman's partial $\rho = -0.1595$, $P = 0.0169$). Minimum F_{ST} showed a significant correlation with both adult and larval τ (adult τ: Spearman's partial $\rho = 0.1703$, $P = 0.01579$; larval τ: Spearman's partial $\rho = 0.1773$, $P = 0.01253$). These results parallel observations made on longer time scales: proteins experience greater adaptive constraint when they have broader functions and more central network positions. Therefore, these factors not only affect substitutions rates through time, but also patterns of polymorphism between different populations.

Regions of adaptive divergence show no effects of genetic constraint

Regions that are involved in local adaptation likely show signs of high K_{ST} or F_{ST}. We therefore repeated the correlations with degree, tissue-bias, and centrality only we represented each gene with the window showing the greatest level of population differentiation. In this instance, we found no significant correlations, suggesting that genetic constraints do not act uniformly across a gene.

Similarly, we looked at the extreme windows for z-scores calculated from other population genetic measurements to see if degree, tissue-bias, and centrality correlated with other

estimates of recent adaptation. As with the K_{ST} and F_{ST} analyses, we chose the window that had the most extreme value across our five populations. We found no significant correlations with minimum Tajima's *D*, however, minimum Fay and Wu's *H* was significantly correlated with betweenness (Spearman's partial $\rho = 0.1822$, $P = 0.006095$), total degree (Spearman's partial $\rho = 0.1678$, $P = 0.01175$), and non-immune degree (Spearman's partial $\rho = 0.1673$, $P = 0.01205$). Negative values of Fay and Wu's *H* indicate the presence of high frequency derived alleles, a signature of a selective sweep. These results therefore suggest that proteins holding central network positions with many interacting partners are less likely to contain adaptively evolving regions.

Here, however, the pattern we observe is driven by differences between proteins that do and do not interact directly with pathogens. When limited to signaling proteins, our analyses no longer showed a positive correlation between Fay and Wu's *H* and degree or Fay and Wu's *H* and betweenness (Spearman's partial rank correlation, $P = 0.2957$ and $P = 0.9654$, respectively). Similarly, we found no correlations when analyzing only effector and recognition proteins (Spearman's partial rank correlation, degree: $P = 0.1116$; betweenness: $P = 0.06038$). As discussed above, betweenness and degree differ among gene classes. In addition, signaling molecules are known to evolve more slowly than recognition and effector proteins over both long and short time-scales (Chapter 3; Sackton, et al. 2007). The correlations we observe may therefore be driven by interactions with the environment, not with other cellular proteins. A similar hypothesis was proposed by Kim, et al. (2007). They noted that proteins at the network periphery were also enriched at the cellular periphery and were therefore subject to increased environmental selection pressures. These results then agree with the population

differentiation analysis above and suggest that genetic constraints do not universally act to limit the progression of adaptation.

Conclusion

As we move towards developing an ever more nuanced understanding of the process of immune adaptation, it will be necessary to incorporate a wider range of parameters into our analyses at all timescales. In studies of short-term adaptation, attention often rests on identifying environmentally imposed selection pressures (Olson-Manning, et al. 2012). It is well-known, however, that non-environmental factors—such as network structure, protein length, gene expression pattern, codon bias, and GC-content—play key roles in promoting or constraining genetic evolution on long time-scales. The patterns we present provide evidence that such genetic constraints also shape the route of local adaptation in *D. melanogaster*. By incorporating this knowledge into an analysis *a priori*, we can set better-informed prior expectations about how adaptation is likely to act across a given set of genes.

Past work has shown that the long-term evolutionary effects of protein-protein interactions are highly context dependent (Mintseris and Weng 2005). Here too, we find that the type of interaction and the strength of selection modulate the outcomes we see. When genes experience strong selection, the effect of genetic constraint is far-reaching and scales with the gene's number of interacting partners. When weaker selection is present, however, recombination may limit the scope of these constraints. This imposes strong conservation on certain key regions, while still allowing normal levels of population differentiation elsewhere in the gene. In short, one metric is not sufficient, but by combining information on gene function, selection

strength, and network structure, we can shed further light on the complex factors that shape the direction and tempo of local adaptation.

Acknowledgements

Betweenness centrality calculations were performed by Yu Guo. We also wish to thank Tim Connallon for helpful discussions.

References

Alvarez-Ponce D, Guirao-Rico S, Orengo DJ, Segarra C, Rozas J, Aguade M 2012. Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in Drosophila melanogaster. Molecular Biology and Evolution 29: 123-132. doi: 10.1093/molbev/msr160

Casals F, Sikora M, Laayouni H, Montanucci L, Muntasell A, Lazarus R, Calafell F, Awadalla P, Netea MG, Bertranpetit J 2011. Genetic adaptation of the antibacterial human innate immunity network. BMC Evol Biol 11: 202. doi: 10.1186/1471-2148-11-202

Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M 2013. The BioGRID interaction database: 2013 update. Nucleic Acids Res 41: D816-823. doi: 10.1093/nar/gks1158

Chintapalli VR, Wang J, Dow JA 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nature Genetics 39: 715-720. doi: 10.1038/ng2049

Clark NL, Alani E, Aquadro CF 2012. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Research 22: 714-720. doi: 10.1101/gr.132647.111

Dupas S, Dubuffet A, Carton Y, Poirie M 2009. Local, geographic and phylogenetic scales of coevolution in Drosophila-parasitoid interactions. Adv Parasitol 70: 281-295. doi: 10.1016/S0065-308X(09)70011-9

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA 2010. Drosophila melanogaster

recombination rate calculator. Gene 463: 18-20. doi: 10.1016/J.Gene.2010.04.015

Fraser HB 2005. Modularity and evolutionary constraint on proteins. Nature Genetics 37: 351-352. doi: 10.1038/ng1530

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW 2002. Evolutionary rate in the protein interaction network. Science 296: 750-752. doi: 10.1126/science.1068696

Hagberg AA, Schult DA, Swart PJ editors. Exploring Network Structure, Dynamics, and Function using NetworkX in Proceedings of the 7th Python in Science conference (SciPy 2008). 2008.

Hahn MW, Kern AD 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution 22: 803-806. doi: 10.1093/molbev/msi072

Hill WG, Zhang XS 2012a. Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. Nature Reviews Genetics 13: 296; author reply 296. doi: 10.1038/nrg2949-c1

Hill WG, Zhang XS 2012b. On the pleiotropic structure of the genotype-phenotype map and the evolvability of complex organisms. Genetics 190: 1131-1137. doi: 10.1534/genetics.111.135681

Kim PM, Korbel JO, Gerstein MB 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc Natl Acad Sci U S A 104: 20274-20279. doi: 10.1073/pnas.0710183104

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG 2008. Evolution of protein-coding genes in Drosophila. Trends Genet 24: 114-123. doi: 10.1016/j.tig.2007.12.001

Loytynoja A, Goldman N 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102: 10557-10562. doi: 10.1073/pnas.0409137102

Luisi P, Alvarez-Ponce D, Dall'Olio GM, Sikora M, Bertranpetit J, Laayouni H 2012. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Molecular Biology and Evolution 29: 1379-1392. doi: 10.1093/molbev/msr298

Mintseris J, Weng Z 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A 102: 10930-10935. doi: 10.1073/pnas.0502667102

Nei M, Gojobori T 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3: 418-426.

Olson-Manning CF, Wagner MR, Mitchell-Olds T 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. Nature Reviews Genetics 13: 867-877. doi: 10.1038/nrg3322

Opulente DA, Morales CM, Carey LB, Rest JS 2013. Coevolution trumps pleiotropy: carbon assimilation traits are independent of metabolic network structure in budding yeast. Plos One 8: e54403. doi: 10.1371/journal.pone.0054403

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG 2007. Dynamic evolution of the innate immune system in Drosophila. Nature Genetics 39: 1461-1468.

doi: 10.1038/ng.2007.60

Vishnoi A, Sethupathy P, Simola D, Plotkin JB, Hannenhalli S 2011. Genome-wide survey of natural selection on functional, structural, and network properties of polymorphic sites in Saccharomyces paradoxus. Molecular Biology and Evolution 28: 2615-2627. doi: 10.1093/molbev/msr085

Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM 2008. Pleiotropic scaling of gene effects and the 'cost of complexity'. Nature 452: 470-472. doi: 10.1038/nature06756

Wagner GP, Zhang J 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nature Reviews Genetics 12: 204-213. doi: 10.1038/nrg2949

Wang Z, Liao BY, Zhang J 2010. Genomic patterns of pleiotropy and the evolution of complexity. Proc Natl Acad Sci U S A 107: 18034-18039. doi: 10.1073/pnas.1004666107

Weir BS, Cockerham CC 1984. Estimating F-Statistics for the Analaysis of Population Structure. Evolution 38: 1358-1370.

APPENDIX 1

Supplementary materials for Chapter 1

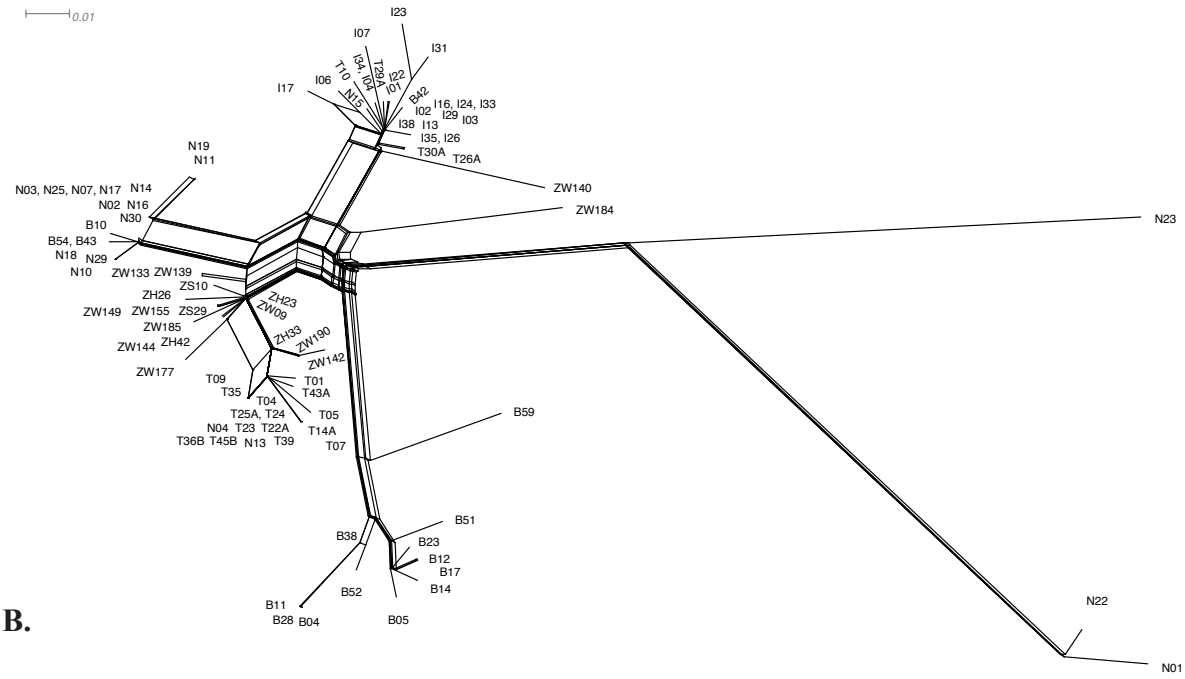
Table S1. Relative levels of commensal bacteria in 37 DGRP inbred lines as measured with qPCR.
Values given for each bacterium are the residuals from a model that accounted for experimental block.

Line	*A. tropicalis*	*L. brevis*	*L. plantarum*	*E. faecalis*
208	-0.860689873	0.709514654	-0.029974816	1.995089889
301	-0.99964351	-0.794096461	-0.763308149	NA
303	1.025912045	1.903125762	1.550774935	-0.534836823
304	-1.260669794	1.510347985	1.021644323	0.675163177
306	3.335742391	3.477570209	0.395580738	-1.264910111
307	-3.071503127	NA	-3.235891725	NA
313	-1.070754621	0.355903542	-0.103863706	1.478398786
315	-0.115199066	1.314236873	-0.068863704	-0.969096178
324	0.600143461	-0.629652015	1.842983638	NA
335	1.892409058	-0.134652015	0.573914073	NA
357	1.002409058	1.090347989	1.151316972	0.025089889
358	1.002578712	1.261459099	0.301691849	NA
360	-2.133532399	-1.271318681	-0.926085927	NA
362	0.793476794	1.623681319	2.324469626	0.627905023
365	0.207578712	-0.540763125	0.003816973	NA
375	-1.172356539	-0.219652016	-0.338683027	5.076717755
379	2.158134268	-1.047985348	1.178914073	0.646717755
380	-0.104003127	-0.176318681	0.167983642	1.465163177
391	0.139800934	0.855903542	2.120025185	-0.644910111
399	2.599245379	0.802014652	0.866316972	-1.424096178
427	-1.53464351	1.203681319	-1.687752593	NA
437	-2.665199066	-4.411318681	-1.44918901	-1.862650534
486	-1.134023206	NA	-0.332522345	-3.739854075
514	0.446467601	1.454792432	-0.332197037	-0.732107912
517	-0.11464351	-0.734652015	-0.119419261	-0.319910111
555	-2.118532399	-2.422152016	-3.184419261	-4.307429513
639	0.073134267	-5.588540904	-4.780349693	-3.237429512
707	-1.351865732	-4.441874237	-5.490349693	-0.780762843
712	-1.305754621	1.372014652	0.262247406	4.685089889
730	1.745976794	0.096459095	2.155580739	2.697905023
732	0.21035649	-2.446874238	-0.912197037	2.405903822
765	2.46035649	2.021181319	2.956136296	3.465236466
774	-0.104087955	0.985903539	0.353358519	-0.499910111
786	1.779800934	2.932014652	1.873358516	1.608403822
799	0.711467601	1.764236875	0.450810988	-2.114910111
820	2.01035649	-0.473818681	-0.821085927	NA
852	-0.412421288	-4.234652016	-0.590385743	NA

APPENDIX 2

Supplementary materials for Chapter 2

A.



B.

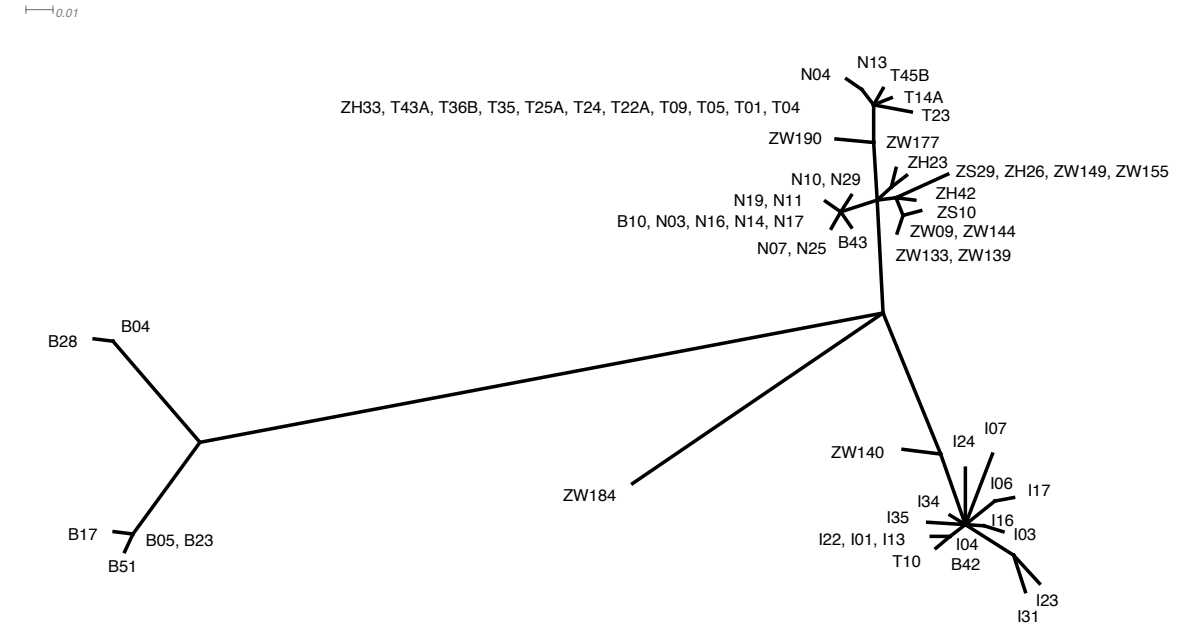


FIGURE S1. Neighbor-Net phylogenetic networks of the (A) 91 mtDNA haplotypes and (B) 65 wMel haplotypes described in Chapter 2.
Networks were constructed with SplitsTree4 v4.12.3 (Huson & Bryant 2006).

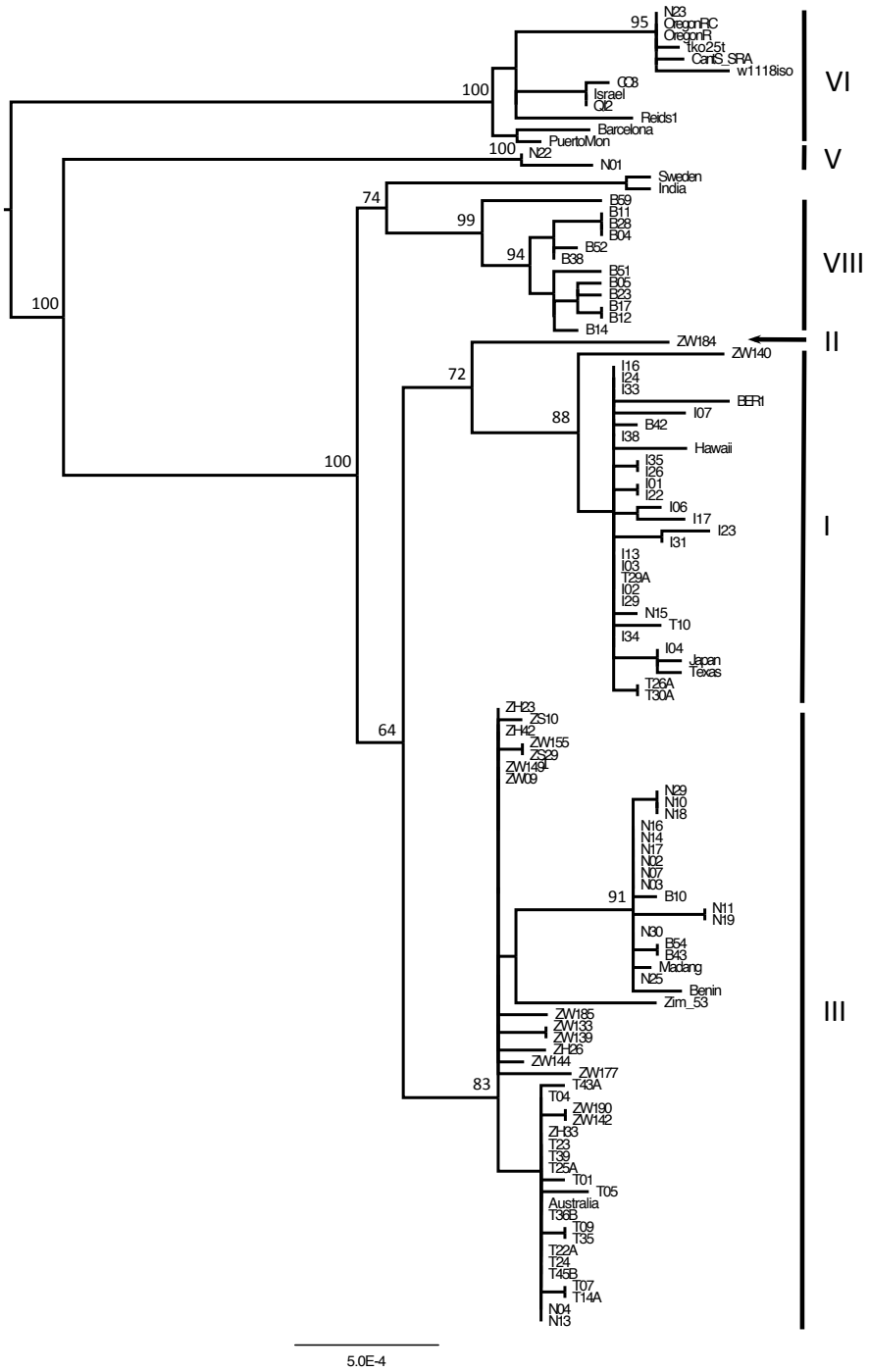


FIGURE S2. Maximum Likelihood tree of D. melanogaster mitochondria sequences. Included are the lines from Chapter 2, 20 additional sequences available on GenBank and a new assembly of short read sequences from a Canton-S line, as described in the main text. RAxML bootstrap values are shown above the nodes and haplotypes are labeled according to Richardson et al. (2012) and Ilinsky (2013). The tree is midpoint rooted and the scale bar is in units of mutations per site.

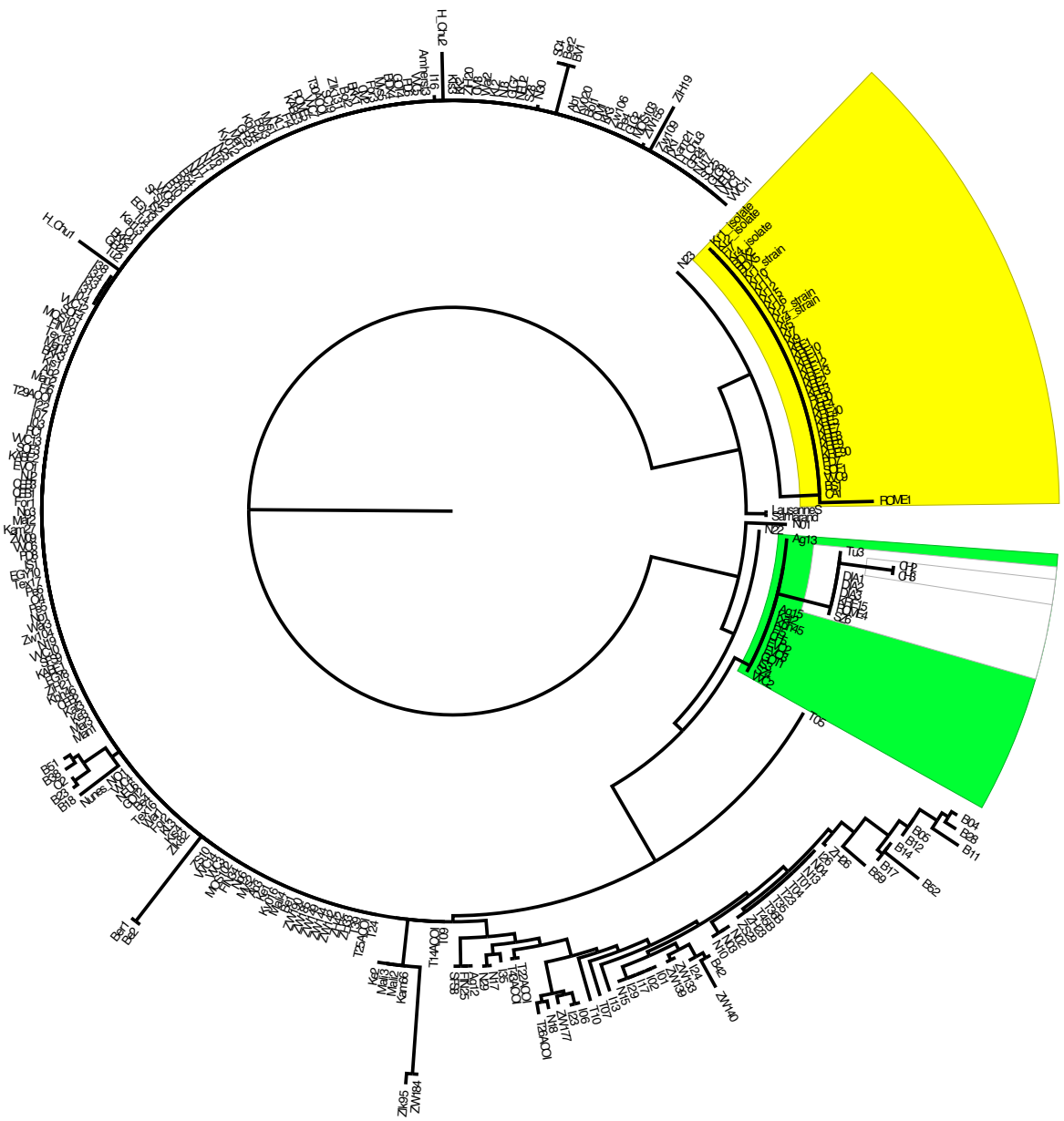


Figure S3. Maximum Likelihood tree of mitochondrial COI sequences from Chapter 2 and Nunes et al. (2008).
Haplotype 1 (yellow) and haplotype 10 (green) are defined by Nunes et al. (2008).

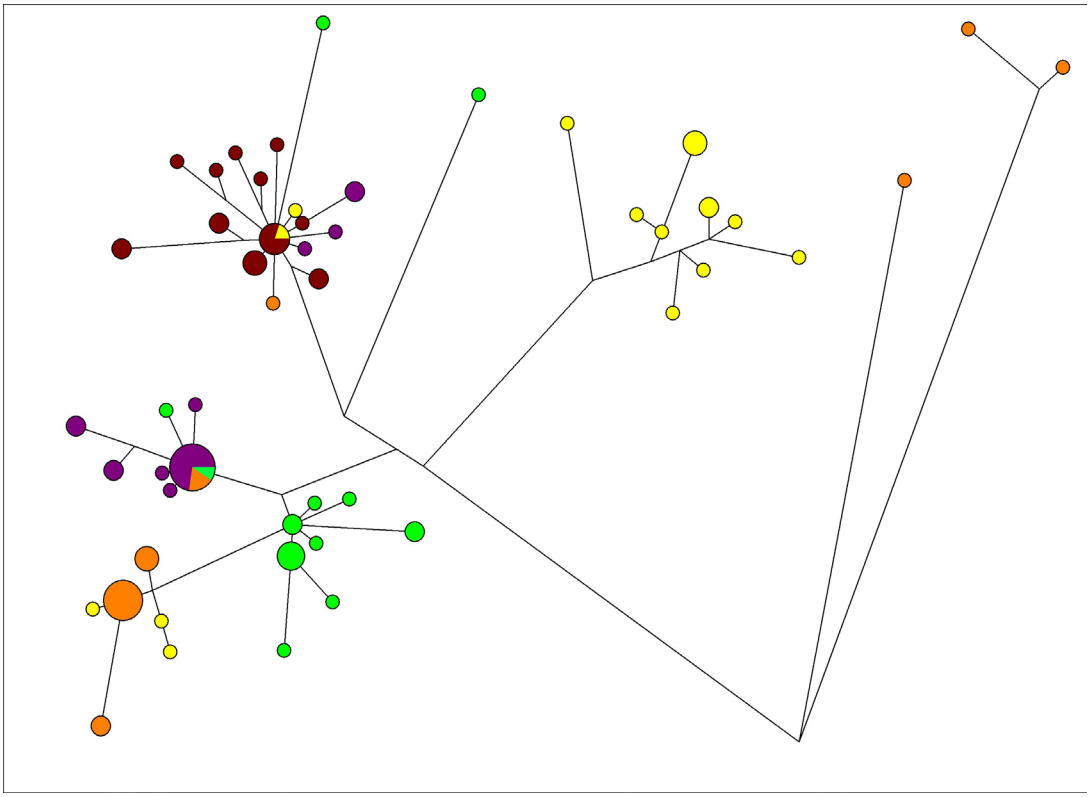


Figure S4. Mitochondrial haplotype network of the 91 *D. melanogaster* lines described in Chapter 2.
The size of each pie is proportional to the number of haplotypes and the colors represent the geographic origin of the fly line: Beijing (yellow), Ithaca, NY (red), Netherlands (orange), Tasmania (purple), Zimbabwe (green).

References

Huson DH, Bryant D 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23: 254-267. doi: 10.1093/molbev/msj030.

Ilinsky Y 2013. Coevolution of Drosophila melanogaster mtDNA and Wolbachia Genotypes. Plos One 8. doi: 10.1371/journal.pone.0054373

Nunes MDS, Nolte V, Schlotterer C 2008b. Nonrandom Wolbachia Infection Status of Drosophila melanogaster Strains with Different mtDNA Haplotypes. Molecular Biology and Evolution 25: 2493-2498. doi: 10.1093/Molbev/Msn199

Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, Bergman CM 2012. Population Genomics of the Wolbachia Endosymbiont in Drosophila melanogaster. Plos Genetics 8: e1003129. doi: 10.1371/journal.pgen.1003129

APPENDIX 3

Supplementary materials for Chapter 3

Table S1. List of D. melanogaster immune genes used in study.
All genes involved in Encapsulation or Phagocytosis were also included in the Cellular category. All genes involved in the IMD or Toll pathways were also included in the Humoral category.

Name	ID	Fly Base ID	Chromosome	Start	End	Functional Class	Immune Process
18 Wheeler	CG8896	FBgn0004364	2R	15999016	16004437	Signaling	Humoral
Adgf-A	CG5992	FBgn0036752	3L	17744831	17757258	Signaling	Cellular
akirin	CG8580	FBgn0082598	3L	7362943	7366811	Signaling	IMD
Alk	CG8250	FBgn0040505	2R	12512798	12527516	Signaling	Encapsulation; PO-Melanization
andropin	CG1361	FBgn0000094	3R	26035670	26036003	Effector (AMP)	IMD
aop	CG3166	FBgn0000097	2L	2156484	2178754	Signaling	JNK; Encapsulation
Aos1	CG12276	FBgn0029512	3R	8258255	8259544	Signaling	Toll
aPKC	CG42783	FBgn0261854	2R	10831963	10850474	Signaling	Toll; Encapsulation
argonaute 2	CG7439	FBgn0087035	3L	15547213	15554142	Anti-viral	Anti-viral
Argonaute-1	CG6671	FBgn0262739	2R	9830892	9845594	Anti-viral	Anti-viral
armitage	CG11513	FBgn0041164	3L	3461317	3466303	Anti-viral	Anti-viral
Ars2	CG7843	FBgn0033062	2R	1968333	1973130	Anti-viral	Anti-viral
Atf-2	CG44246	FBgn0265193	2R	20757631	20759032	Signaling	Epithelial; ROS
Atf3	CG11405	FBgn0028550	X	1136773	1168682	Signaling	Humoral; Epithelial
atilla	CG6579	FBgn0032422	2L	12175451	12177886	Unknown	Encapsulation
att A	CG10146	FBgn0012042	2R	10634867	10635720	Effector (AMP)	IMD
att B1	CG18372	FBgn0041581	2R	10636728	10637670	Effector (AMP)	IMD
att D	CG7629	FBgn0038530	3R	13450990	13451808	Effector (AMP)	IMD
AttC	CG4740	FBgn0041579	2R	9281210	9282172	Effector (AMP)	IMD
aubergine	CG6137	FBgn0000146	2L	10997819	11001476	Anti-viral	Anti-viral
Bag of Marbles/BAM	CG10422	FBgn0000158	3R	21068761	21070779	Signaling	Cellular
basket	CG5680	FBgn0000229	2L	10247502	10250501	Signaling	JNK; Wound repair
Bendless/Ubc13	CG18319	FBgn0000173	X	13890378	13893838	Signaling	IMD
brm	CG5942	FBgn0000212	3L	15963182	15975969	Signaling	Cellular
cactus	CG5848	FBgn0000250	2L	16313036	16326207	Signaling	Humoral; Phagocytosis
Cad96Ca/Stitcher	CG10244	FBgn0022800	3R	21023293	21034005	Signaling	Wound repair
CanB	CG4209	FBgn0010014	X	5226532	5228482	Signaling	Epithelial; ROS
car	CG12230	FBgn0000257	X	19460052	19463434	Other	Phagocytosis
Caspar	CG8400	FBgn0034068	2R	11912133	11915666	Signaling	IMD
Catsup	CG10449	FBgn0002022	2L	19041862	19043614	Signaling	PO-Melanization
Cdc42	CG12530	FBgn0010341	X	19591117	19593971	Signaling	Encapsulation; Phagocytosis; Wound repair
Cdk5	CG8203	FBgn0013762	2R	11457033	11458922	Signaling	Encapsulation
Cdk5a	CG5387	FBgn0027491	2L	10308131	10310956	Signaling	Encapsulation
cecA1	CG1365	FBgn0000276	3R	26036596	26036995	Effector (AMP)	IMD
cecA2	CG1367	FBgn0000277	3R	26037878	26038290	Effector (AMP)	IMD
cecB	CG1878	FBgn0000278	3R	26039037	26039519	Effector (AMP)	IMD
cecC	CG1373	FBgn0000279	3R	26042206	26042660	Effector (AMP)	IMD
CG10345	CG10345	FBgn0027562	3R	12422795	12430583	Recognition	Cellular
CG11159	CG11159	FBgn0034539	2R	16543261	16544114	Effector (Lysozyme)	Other/Unknown
CG12780	CG12780	FBgn0033301	2R	4439364	4439758	Recognition	JAK-STAT; Anti-viral
CG13422	CG13422	FBgn0034511	2R	16413800	16414330	Recognition	Humoral
CG13551	CG13551	FBgn0040660	2R	19265389	19270947	Effector (AMP)	Humoral
CG16756	CG16756	FBgn0029765	X	5321550	5322250	Effector (Lysozyme)	Humoral
CG16799	CG16799	FBgn0034538	2R	16540062	16542601	Effector (Lysozyme)	Humoral
CG18107	CG18107	FBgn0034330	2R	14272072	14272416	Effector (Putative)	Humoral
CG2736	CG2736	FBgn0035090	2R	20860954	20862864	Recognition	Cellular
CG3829	CG3829	FBgn0035091	2R	20873567	20878012	Recognition	Cellular
CG4572	CG4572	FBgn0038738	3R	15688239	15690115	Anti-viral	Anti-viral
CG6426	CG6426	FBgn0034162	2R	12756907	12759356	Effector (Lysozyme)	Humoral
CG6429	CG6429	FBgn0046999	2R	12760520	12761260	Effector (Lysozyme)	Humoral
CG7158	CG7158	FBgn0037116	3L	21667494	21672855	Signaling	Encapsulation
CG7227	CG7227	FBgn0031970	2L	7994388	7998151	Recognition	Phagocytosis
CG7798	CG7798	FBgn0034092	2R	12109617	12110132	Effector (Lysozyme)	Humoral
CG8492	CG8492	FBgn0035813	3L	7720904	7725291	Effector (Lysozyme)	Humoral
cher	CG3937	FBgn0014141	3R	12917255	12951111	Signaling	Cellular
CHKov1	CG10618	FBgn0045761	3R	21148878	21155038	Anti-viral	Anti-viral
CHKov2	CG10675	FBgn0039328	3R	21155252	21156925	Anti-viral	Anti-viral
CHMP2B	CG4618	FBgn0035589	3L	5132719	5133573	Signaling	Toll
Corin	CG2105	FBgn0033192	2R	3427616	3451733	Signaling	Cellular; Coagulation
croquemort	CG4280	FBgn0015924	2L	448254	453024	Recognition	Phagocytosis
daw	CG16987	FBgn0031461	2L	2805261	2812376	Signaling	Toll; PO-Melanization
Ddc	CG10697	FBgn0000422	2L	19116483	19120306	Effector	Wound repair; PO-Melanization; ROS
Deaf1	CG8567	FBgn0013799	3L	19811274	19823786	Signaling	Toll
defensin	CG1385	FBgn0010385	2R	5941683	5942081	Effector (AMP)	Toll
dFADD/BG4	CG12297	FBgn0038928	3R	17860177	17861339	Signaling	IMD
Dhc64C	CG7507	FBgn0261797	3L	4807368	4825687	Signaling	Encapsulation
dia	CG1768	FBgn0011202	2L	20758144	20768053	Signaling	Encapsulation
Dicer-1	CG4792	FBgn0039016	3R	18559729	18566813	Anti-viral	Anti-viral
Dicer-2	CG6493	FBgn0034246	2R	13462484	13469030	Anti-viral	Anti-viral
Diedel	CG11501	FBgn0039666	3R	25319574	25319989	Signaling	JAK-STAT; Anti-viral
DIF	CG6794	FBgn0011274	2L	17413248	17433161	Signaling	IMD; Toll
diptericin	CG12763	FBgn0004240	2R	14753270	14753765	Effector (AMP)	IMD
dnr1	CG12489	FBgn0260866	2R	18450657	18480469	Signaling	IMD
dom	CG9696	FBgn0020306	2R	17210935	17229352	Signaling	IMD; Cellular
Domeless	CG14226	FBgn0043903	X	19570094	19577551	Signaling	JAK-STAT
dorsal	CG6667	FBgn0260632	2L	17436830	17450364	Signaling	Toll; PO-Melanization
dpp	CG9885	FBgn0000490	2L	2428372	2459823	Signaling	Humoral; Cellular
dpt B	CG10794	FBgn0034407	2R	14754896	14755400	Effector (AMP)	IMD
Draper	CG2086	FBgn0027594	3L	1715595	1731107	Recognition	Phagocytosis
dredd	CG7486	FBgn0020381	X	527655	529578	Signaling	IMD; JNK
dro2	CG32279	FBgn0052279	3L	3314349	3314681	Effector (AMP)	Toll
dro3	CG32283	FBgn0052283	3L	3314996	3315355	Effector (AMP)	Toll
dro4	CG32282	FBgn0052282	3L	3315620	3315942	Effector (AMP)	Toll
dro5	CG10812	FBgn0035434	3L	3316781	3317144	Effector (AMP)	Toll

TABLE S1 (continued).

Name	ID	Fly Base ID	Chromosome	Start	End	Functional Class	Immune Process
dro6	CG32268	FBgn0052268	3L	3336032	3336418	Effector (AMP)	Toll
drosocin	CG10816	FBgn0010388	2R	10633466	10634219	Effector (AMP)	IMD
drs	CG10810	FBgn0010381	3L	3369556	3369942	Effector (AMP)	Toll
Drs-l	CG32274	FBgn0052274	3L	3335579	3335788	Effector (AMP)	Toll
Dscam	CG17800	FBgn0033159	2R	3205429	3269404	Recognition	Phagocytosis
Dsor1	CG15793	FBgn0010269	X	9141375	9144074	Signaling	Cellular
Dsp1	CG12223	FBgn0011764	X	16227645	16234657	Signaling	Humoral
duox	CG3131	FBgn0031464	2L	2815970	2830248	Effector (ROS)	Epithelial; ROS
ea	CG4920	FBgn0000533	3R	11154451	11156130	Signaling	Humoral
Eater	CG6124	FBgn0243514	3R	22921589	22925401	Recognition	Phagocytosis; Anti-viral
Eb1	CG3265	FBgn0027066	2R	2636732	2643805	Signaling	Encapsulation
ecd	CG5714	FBgn0000543	3L	2263581	2265857	Signaling	Encapsulation
ECSIT	CG10610	FBgn0028436	3R	1682307	1683636	Signaling	Toll
Ect4	CG43119	FBgn0262579	3L	8056974	8101936	Signaling	Epithelial
edin	CG32185	FBgn0052185	3L	17487980	17488400	Signaling	IMD
edl	CG15085	FBgn0023214	2R	14555032	14561041	Signaling	Cellular
effete	CG7425	FBgn0011217	3R	10558136	10567041	Signaling	IMD
Egfr	CG10079	FBgn0003731	2R	17409925	17447482	Signaling	Cellular
egghead	CG9659	FBgn0001404	X	2482536	2493144	Anti-viral	Anti-viral
Eig71Ee	CG7604	FBgn0004592	3L	15647572	15649110	Effector	Coagulation
eiger	CG12919	FBgn0033483	2R	5965871	5971715	Signaling	JNK; PO-Melanization
emb	CG13387	FBgn0020497	2L	8403573	8408853	Signaling	Toll
emp	CG2727	FBgn0010435	2R	20863979	20872321	Recognition	Cellular
enok	CG11290	FBgn0034975	2R	19986733	19994555	Signaling	JAK-STAT
Eph	CG1511	FBgn0025936	4	631310	643096	Signaling	Encapsulation
Ephrin	CG1862	FBgn0040324	4	592381	600748	Signaling	Encapsulation
Exn	CG42665	FBgn0261547	3L	16963113	16987488	Signaling	Encapsulation
eye transformer	CG14225	FBgn0031055	X	19566345	19569924	Signaling	JAK-STAT; Encapsulation
Fer2LCH	CG1469	FBgn0015221	3R	26213526	26216306	Effector	Other/Unknown
Flotillin2	CG32593	FBgn0264078	X	14733409	14827979	Signaling	Encapsulation
Fondue	CG15825	FBgn0032773	2L	19383138	19386205	Effector	Coagulation
G protein alpha49B	CG17759	FBgn0004435	2R	8500245	8510702	Signaling	Epithelial; ROS
Gadd45	CG11086	FBgn0033153	2R	3136666	3138160	Signaling	JNK
GATAe	CG10278	FBgn0038391	3R	11834827	11844855	Signaling	Epithelial
gcm2	CG3858	FBgn0019809	2L	9608562	9612706	Signaling	Cellular; PO-Melanization
Ggamma1	CG8261	FBgn0004921	2R	4788909	4792296	Other	Wound repair
glial cells missing	CG12245	FBgn0014179	2L	9579498	9581745	Signaling	Cellular; PO-Melanization
Glued	CG9206	FBgn0001108	3L	13922487	13927756	Signaling	Encapsulation
GNBP1	CG6895	FBgn0040323	3L	18668900	18670976	Recognition	Toll
GNBP2	CG4144	FBgn0040322	3L	18666574	18668513	Recognition	Humoral
GNBP3	CG5008	FBgn0040321	3L	8948121	8949651	Recognition	Toll
Gp150	CG5820	FBgn0013272	2R	18205167	18216921	Recognition	Cellular
Gprk2	CG17998	FBgn0261988	3R	27230967	27283595	Signaling	Toll
Grainy head	CG42311	FBgn0259211	2R	13689751	13730325	Signaling	Wound repair
Grass	CG5896	FBgn0039494	3R	22983670	22985503	Signaling	Toll
grim	CG4345	FBgn0015946	3L	18295819	18297517	Signaling	PO-Melanization
Gustatory receptor 28b	CG13788	FBgn0045495	2L	7454826	7462247	Signaling	PO-Melanization
Hayan	CG6361	FBgn0030925	X	18374016	18377805	Signaling	Wound repair; PO-Melanization
Helicase89B	CG4261	FBgn0022787	3R	11799020	11809173	Signaling	IMD; Toll
Hemese	CG31770	FBgn0028430	2L	13972096	13972947	Recognition	Encapsulation
hemipterous/hep	CG4353	FBgn0010303	X	12973238	12984467	Signaling	JNK; Wound repair; PO-Melanization
Hml	CG7002	FBgn0029167	3L	13839154	13853101	Effector	Cellular; Coagulation; Wound repair
hopscotch/Jak	CG1594	FBgn0004864	X	11254963	11262131	Signaling	JAK-STAT; Anti-viral
Hrs	CG2903	FBgn0031450	2L	2739986	2743377	Signaling	Toll; Cellular
Iap2	CG8293	FBgn0015247	2R	11819675	11822330	Signaling	IMD
IM1	CG18108	FBgn0034329	2R	14271454	14271883	Effector (Putative)	Toll
IM10	CG18279	FBgn0033835	2R	9294516	9295863	Effector (Putative)	Humoral
IM2	CG18106	FBgn0025583	2R	14274107	14274535	Effector (Putative)	Toll
IM23	CG15066	FBgn0034328	2R	14270209	14270740	Effector (Putative)	Toll
IM3	CG16844	FBgn0040736	2R	14275400	14276239	Effector (Putative)	Toll
IM4	CG15231	FBgn0040653	2R	16756350	16756826	Effector (Putative)	Humoral
imd	CG5576	FBgn0013983	2R	14297301	14299037	Signaling	IMD
IRC	CG8913	FBgn0038465	3R	12828597	12832930	Effector (ROS)	Epithelial; ROS
ird5/IKKbeta	CG4201	FBgn0024222	3R	11872069	11875143	Signaling	IMD
itgBv	CG1762	FBgn0010395	2L	21053033	21058044	Recognition	Encapsulation; Phagocytosis
Jafrac1	CG1633	FBgn0040309	X	13223857	13226027	Effector (ROS)	ROS
Jafrac2	CG1274	FBgn0040308	3L	3042359	3044174	Effector (ROS)	ROS
Jra/AP1	CG2275	FBgn0001291	2R	5983985	5986061	Signaling	IMD; JNK; Phagocytosis; Wound repair
kay	CG33956	FBgn0001297	3R	25592497	25619838	Signaling	IMD; JNK; Wound repair
Keap1	CG3962	FBgn0038475	3R	12899664	12905667	Signaling	Other/Unknown
ken	CG5575	FBgn0011236	2R	19757796	19764965	Signaling	JAK-STAT
kenny	CG16910	FBgn0041205	2R	20673037	20674934	Signaling	IMD; Toll; Anti-viral
kkv	CG2666	FBgn0001311	3R	1203815	1218588	Effector	Wound repair
knot	CG10197	FBgn0001319	2R	10660155	10694566	Signaling	Cellular
l(3)mbn	CG12755	FBgn0002440	3L	6117034	6121907	Signaling	Cellular
LanA	CG10236	FBgn0002526	3L	6196931	6211122	Recognition	Encapsulation
lectin-24A	CG3410	FBgn0040104	2L	3716800	3717773	Recognition	Cellular
lectin-37Da	CG33532	FBgn0053532	2L	19418317	19419029	Recognition	Encapsulation
lectin-37Db	CG33533	FBgn0053533	2L	19419075	19419767	Recognition	Encapsulation
lectin-galC1	CG9976	FBgn0016675	2L	19417447	19418274	Recognition	Cellular
licorne/Mkk3	CG12244	FBgn0261524	X	12985178	12988019	Signaling	Epithelial; ROS
Listericin	CG9080	FBgn0033593	2R	7129250	7129699	Effector (AMP)	IMD; JAK-STAT; Anti-viral
lola	CG12052	FBgn0005630	2R	6369399	6430796	Signaling	IMD; Toll; Phagocyosis
lozenge/lz	CG1689	FBgn0002576	X	9178682	9197669	Signaling	Cellular; Wound repair; PO-Melanization

TABLE S1 (continued).

Name	ID	Fly Base ID	Chromosome	Start	End	Functional Class	Immune Process
LpR1	CG31094	FBgn0066101	3R	21567070	21588827	Signaling	Humoral
lwr/Ubc9	CG3018	FBgn0010602	2L	540599	542580	Signaling	Toll; Cellular
lysB	CG1179	FBgn0004425	3L	1207371	1207981	Effector (Lysozyme)	Other/Unknown
lysC	CG9111	FBgn0004426	3L	1210485	1210971	Effector (Lysozyme)	Other/Unknown
lysD	CG9118	FBgn0004427	3L	1210717	1211196	Effector (Lysozyme)	Other/Unknown
lysE	CG1180	FBgn0004428	3L	1212827	1213406	Effector (Lysozyme)	Other/Unknown
lysP	CG9116	FBgn0004429	3L	1218082	1218696	Effector (Lysozyme)	Other/Unknown
lysS	CG1165	FBgn0004430	3L	1227747	1228238	Effector (Lysozyme)	Other/Unknown
lysX	CG9120	FBgn0004431	3L	1194739	1195302	Effector (Lysozyme)	Other/Unknown
mask	CG33106	FBgn0043884	3R	20056284	20075970	Signaling	IMD; Cellular
mbc	CG10379	FBgn0015513	3R	19607487	19627356	Other	Wound repair
mbo	CG6819	FBgn0026207	3R	8502209	8504774	Signaling	IMD
Mcr	CG7586	FBgn0264800	2L	8074117	8082982	Recognition	Phagocytosis
Mekk1	CG7717	FBgn0024329	3R	14558828	14569316	Signaling	Other/Unknown
metchnikowin	CG8175	FBgn0014865	2R	11296351	11296618	Effector (AMP)	Toll
Mkk4	CG9738	FBgn0024326	3R	4471894	4476518	Signaling	JNK; Wound repair
Mkp3	CG14080	FBgn0036844	3L	19059255	19079526	Signaling	Epithelial; ROS
Mmp1	CG4859	FBgn0035049	2R	20558817	20575707	Signaling	JNK; Wound repair
modSP	CG31217	FBgn0051217	3R	12478168	12481384	Signaling	Toll
MP1	CG1102	FBgn0027930	3R	133630	135557	Signaling	PO-Melanization; Epithelial
Mpk2/p38a	CG5475	FBgn0015765	3R	19976138	19978043	Signaling	Other/Unknown
msn	CG16973	FBgn0010909	3L	2554847	2586540	Effector	JNK; Wound repair
mtd	CG32464	FBgn0013576	3R	1095838	1176165	Signaling	IMD
mxc/multi sex combs	CG12124	FBgn0260789	X	9130834	9137500	Signaling	Cellular
Myd88	CG2078	FBgn0033402	2R	5190328	5196227	Signaling	Toll
myopic	CG9311	FBgn0036448	3L	14762441	14769229	Signaling	Toll
myospheroid	CG1560	FBgn0004657	X	7955678	7964270	Recognition	Encapsulation
Nec/spn43Ac	CG1857	FBgn0002930	2R	3043982	3045746	Signaling	Toll
Neuroglian	CG1634	FBgn0264975	X	8411406	8449203	Recognition	Encapsulation
nimA	CG42282	FBgn0261514	2L	13957539	13963381	Recognition	Phagocytosis
nimB1	CG33119	FBgn0027929	2L	13963507	13965088	Recognition	Phagocytosis
nimB2	CG31839	FBgn0028543	2L	13965012	13967018	Recognition	Phagocytosis
nimB3	CG34003	FBgn0054003	2L	13967464	13967985	Recognition	Phagocytosis
nimB4	CG33115	FBgn0028542	2L	13968280	13969976	Recognition	Phagocytosis
nimB5	CG16873	FBgn0028936	2L	13970177	13972173	Recognition	Phagocytosis
nimC1	CG8942	FBgn0259896	2L	13973158	13976769	Recognition	Phagocytosis
nimC2	CG18146	FBgn0028939	2L	13980126	13983269	Recognition	Phagocytosis
nimC3	CG16880	FBgn0001967	2L	14019357	14020208	Recognition	Phagocytosis
nitric oxide synthase	CG6713	FBgn0011676	2L	10804274	10837511	Signaling	IMD; ROS
norpA	CG3620	FBgn0262738	X	4216659	4259417	Signaling	Epithelial; ROS
Not4	CG31716	FBgn0051716	2L	10396716	10403206	Signaling	JAK-STAT
Notch	CG3936	FBgn0004647	X	3028903	3066254	Signaling	JNK; JAK-STAT; Cellular; PO-Melanization
Npc2g	CG11314	FBgn0039800	3R	26570327	26570985	Effector	Coagulation; Wound repair
Npc2h	CG11315	FBgn0039801	3R	26571201	26572012	Effector	Coagulation; Wound repair
NT1/Spz2	CG42576	FBgn0261526	3L	4501652	4508433	Signaling	Epithelial
Ntf-2	CG1740	FBgn0031145	X	20906496	20912494	Signaling	Humoral
Nubbin	CG34395	FBgn0085424	2L	12587625	12628135	Signaling	Epithelial
Nup214	CG3820	FBgn0010660	2R	18806460	18812265	Signaling	Toll
NURF	CG32346	FBgn0000541	3L	233926	246912	Signaling	JAK-STAT; Cellular
os/upd1	CG5993	FBgn0004956	X	18199387	18203251	Signaling	JAK-STAT
p38b	CG7393	FBgn0024846	2L	13780689	13782611	Signaling	Other/Unknown
p38c	CG33338	FBgn0046322	3R	19974742	19975925	Signaling	PO-Melanization; ROS
p53	CG33336	FBgn0039044	3R	18875379	18879804	Anti-viral	Anti-viral
Pale	CG10118	FBgn0005626	3L	6707138	6712625	Effector	Wound repair; PO-Melanization
pannier	CG3978	FBgn0003117	3R	11851921	11867526	Signaling	Toll
pastrel	CG8588	FBgn0035770	3L	7350375	7353363	Anti-viral	Anti-viral
Pebp1	CG18594	FBgn0038973	3R	18291756	18292401	Signaling	JAK-STAT; Coagulation; PO-Melanization
pelle	CG5974	FBgn0010441	3R	23076853	23078904	Signaling	Toll
pellino	CG5212	FBgn0025574	3R	19685884	19715774	Signaling	Toll
persephone	CG6367	FBgn0030926	X	18378520	18380954	Signaling	Toll
peste	CG7228	FBgn0031969	2L	7986786	7994184	Recognition	Phagocytosis
PGRP-LA	CG32042	FBgn0035975	3L	9327432	9331436	Recognition	IMD; Epithelial
PGRP-LB	CG14704	FBgn0037906	3R	7278571	7286274	Recognition	IMD; Epithelial
PGRP-LC	CG4432	FBgn0035976	3L	9331910	9341436	Recognition	IMD; Phagocytosis; Epithelial
PGRP-LD	CG33717	FBgn0260458	3L	5773151	5777785	Recognition	IMD; Epithelial
PGRP-LE	CG8995	FBgn0030695	X	15695186	15697798	Recognition	IMD; Epithelial
PGRP-LF	CG4437	FBgn0035977	3L	9342709	9344589	Recognition	IMD; Epithelial
PGRP-SA	CG11709	FBgn0030310	X	11455676	11456812	Recognition	Toll
PGRP-SB1	CG9681	FBgn0043578	3L	16720399	16721089	Recognition	Humoral
PGRP-SB2	CG9697	FBgn0043577	3L	16719641	16720388	Recognition	Humoral
PGRP-SC1a	CG14746	FBgn0043576	2R	4597238	4597825	Recognition	IMD; Phagocytosis; Epithelial
PGRP-SC1b	CG8577	FBgn0033327	2R	4600951	4602599	Recognition	IMD; Epithelial
PGRP-SC2	CG14745	FBgn0043575	2R	4604455	4605200	Recognition	IMD
PGRP-SD	CG7496	FBgn0035806	3L	7644280	7645000	Recognition	Toll
phl	CG2845	FBgn0003079	X	2189499	2237903	Signaling	Cellular
pirk	CG15678	FBgn0034647	2R	17548472	17549772	Signaling	IMD; Epithelial
piwi	CG6122	FBgn0004872	2L	10982205	10987420	Anti-viral	Anti-viral
pnt	CG17077	FBgn0003118	3R	19115953	19171889	Signaling	Cellular
poly	CG9829	FBgn0086371	3R	9187538	9192632	Signaling	Encapsulation
POSH	CG4909	FBgn0040294	2R	13456237	13459827	Signaling	IMD; JNK
proPO-A1	CG5779	FBgn0261362	2R	13774725	13777478	Effector	Wound repair; PO-Melanization
proPO45	CG8193	FBgn0033367	2R	4929766	4932214	Effector	Coagulation; Wound repair; PO-Melanization
proPO59	CG42640	FBgn0261363	2R	18952292	18954706	Effector	Encapsulation; PO-Melanization
Prx5	CG7217	FBgn0038570	3R	13991164	13992995	Effector	ROS

TABLE S1 (continued).

Name	ID	Fly Base ID	Chromosome	Start	End	Functional Class	Immune Process
psidin	CG4845	FBgn0243511	3R	15847426	15852204	Signaling	Humoral; Phagocytosis
Pten	CG5671	FBgn0026379	2L	10256319	10261049	Signaling	Encapsulation
Puckered	CG7850	FBgn0243512	3R	3931054	3948016	Signaling	JNK; Wound repair
Punch	CG9441	FBgn0003162	2R	17062820	17070113	Signaling	IMD; JNK; PO-Melanization
Putzig	CG7752	FBgn0259785	3L	21279199	21283410	Signaling	JAK-STAT
Pvf1	CG7103	FBgn0030964	X	18726715	18736934	Signaling	Cellular; Wound repair
Pvf2	CG13780	FBgn0031888	2L	7069536	7084635	Signaling	Cellular
Pvf3	CG34378	FBgn0085407	2L	7098188	7157695	Signaling	Cellular
Pvr	CG8222	FBgn0032006	2L	8220980	8239878	Signaling	IMD; Cellular; Wound repair
r2d2	CG7138	FBgn0031951	2L	7800147	7802098	Anti-viral	Anti-viral
Rab7	CG5915	FBgn0015795	3R	19818730	19821174	Effector	Phagocytosis
Rac1	CG2248	FBgn0010333	3L	1300879	1302683	Signaling	JNK; Encapsulation; Phagocytosis; Wound repair
Rac2	CG8556	FBgn0014011	3L	7426749	7428290	Signaling	IMD; JNK; Encapsulation; Phagocytosis; Wound
Rap1	CG1956	FBgn0004636	3L	1859115	1862239	Signaling	Cellular
Ras85D	CG9375	FBgn0003205	3R	5336283	5338789	Signaling	Cellular
ref(2)P	CG10360	FBgn0003231	2L	19542468	19545548	Signaling	Toll; Anti-viral
relish	CG11992	FBgn0014018	3R	4869905	4873699	Signaling	IMD
RfaBp	CG11064	FBgn0087002	4	1085532	1097899	Signaling	Coagulation
Rho1	CG8416	FBgn0014020	2R	11990192	11994734	Signaling	Cellular; Wound repair; PO-Melanization
RhoBTB	CG5701	FBgn0036980	3L	20369342	20375498	Signaling	Encapsulation
RhoGEF3	CG43976	FBgn0264707	3L	277437	305292	Signaling	Encapsulation
RhoL	CG9366	FBgn0014380	3R	5322823	5328303	Signaling	Encapsulation
rl/rolled/ERK	CG12559	FBgn0003256	2RHet	198176	251950	Signaling	Wound repair
Rm62	CG10279	FBgn0003261	3R	1826148	1834301	Anti-viral	Anti-viral
Rpr	CG4319	FBgn0011706	3L	18390635	18391535	Signaling	IMD
santa-maria	CG12789	FBgn0025697	2L	7445355	7451167	Recognition	Cellular
scb	CG8095	FBgn0003328	2R	11136289	11146003	Recognition	Phagocytosis
sec5	CG8843	FBgn0266670	2L	3457302	3460367	Anti-viral	Anti-viral
serpent	CG3992	FBgn0003507	3R	11811874	11829803	Signaling	Cellular; PO-Melanization
Serpin 27A	CG11331	FBgn0028990	2L	6671065	6673347	Signaling	PO-Melanization
Serpin 28D	CG7219	FBgn0031973	2L	8004548	8007210	Signaling	PO-Melanization
Serrate	CG6127	FBgn0004197	3R	22997812	23019989	Signaling	Cellular; PO-Melanization
Shc	CG3715	FBgn0015296	3L	9420110	9421771	Signaling	Encapsulation
shrub	CG8055	FBgn0086656	2R	5029651	5031419	Other	Phagocytosis
sick	CG42589	FBgn0263873	2L	19796365	19958424	Signaling	IMD
singed/fascin	CG32858	FBgn0003447	X	7858057	7880634	Signaling	Cellular
slpr	CG2272	FBgn0030018	X	8064408	8071218	Signaling	JNK; Wound repair
smt3	CG4494	FBgn0264922	2L	6966780	6967644	Signaling	IMD
Socs36E	CG15154	FBgn0041184	2L	18138675	18152417	Signaling	JAK-STAT
Socs44A	CG2160	FBgn0033266	2R	4015080	4016981	Signaling	JAK-STAT
Sp7	CG3066	FBgn0037515	3R	3585508	3589295	Signaling	PO-Melanization; Epithelial
spatzle	CG6134	FBgn0003495	3R	22890712	22895792	Signaling	Toll
SPE	CG16705	FBgn0039102	3R	19511977	19513869	Signaling	Toll
spheroide	CG9675	FBgn0030774	X	16594479	16596005	Signaling	Toll
sphinx1	CG32383	FBgn0052383	3L	7431521	7432505	Signaling	Toll
sphinx2	CG32382	FBgn0052382	3L	7432799	7433833	Signaling	Toll
Spirit	CG2056	FBgn0030051	X	8465132	8467626	Signaling	Toll
spn-E	CG3158	FBgn0003483	3R	11663141	11670137	Anti-viral	Anti-viral
Spn1	CG9456	FBgn0028988	2R	2770785	2772378	Signaling	Toll
spn5	CG18525	FBgn0028984	3R	11028791	11032169	Signaling	Toll; PO-Melanization
Spn77Ba	CG6680	FBgn0262057	3L	20310290	20313047	Signaling	PO-Melanization; Epithelial
Sr-CI	CG4099	FBgn0014033	2L	4121702	4124006	Recognition	Phagocytosis; Anti-viral
Sr-CII	CG8856	FBgn0020377	2R	8096398	8099660	Recognition	Phagocytosis
Sr-CIII	CG31962	FBgn0020376	2L	4120343	4121627	Recognition	Phagocytosis
Sr-CIV	CG3212	FBgn0031547	2L	3522521	3524056	Recognition	Phagocytosis
Stam	CG6521	FBgn0027363	2L	10858536	10862006	Signaling	Cellular; Wound repair
STAT92E	CG4257	FBgn0016917	3R	16361045	16378033	Signaling	Humoral; JAK-STAT; Epithelial
Su(H)	CG3497	FBgn0004837	2L	15039488	15043334	Signaling	Cellular; PO-Melanization
Su(var)2-10	CG8068	FBgn0003612	2R	5003627	5009505	Signaling	JAK-STAT
Tab2	CG7417	FBgn0086358	2R	15180034	15192013	Signaling	IMD; JNK
Taf1	CG17603	FBgn0010355	3R	2472611	2481770	Signaling	JAK-STAT
Tak1	CG18492	FBgn0026323	X	20386935	20395955	Signaling	IMD; JNK; Wound repair; Anti-viral
Takl2	CG4803	FBgn0039015	3R	18558078	18559573	Signaling	Wound repair
tamo	CG4057	FBgn0041582	2R	20014161	20021074	Signaling	Toll
TepI	CG18096	FBgn0041183	2L	15888640	15893811	Recognition	Toll; JAK-STAT; Phagocytosis
TepII	CG7052	FBgn0041182	2L	7693727	7701601	Recognition	Phagocytosis
TepIII	CG7068	FBgn0041181	2L	7702805	7710690	Recognition	Phagocytosis
TepIV	CG10363	FBgn0041180	2L	19548507	19556451	Recognition	Humoral; Cellular
Tetraspanin 68C	CG32136	FBgn0043550	3L	13836756	13838416	Signaling	Cellular
Thor	CG8846	FBgn0261560	2L	3478434	3479612	Signaling	Humoral; Phagocytosis
thread/Iap1	CG12284	FBgn0260635	3L	16031510	16044134	Signaling	Cellular
Tig	CG11527	FBgn0011722	2L	6415179	6423310	Other	Cellular; Coagulation
Tl-3/MstProx	CG1149	FBgn0015770	3R	3191661	3195027	Signaling	Other/Unknown
TM9SF4	CG7364	FBgn0028541	2L	13746952	13773905	Recognition	Encapsulation; Phagocytosis; PO-Melanization
Toll	CG5490	FBgn0262473	3R	22624763	22668125	Signaling	Toll
Toll-4	CG18241	FBgn0032095	2L	9084107	9089440	Signaling	Other/Unknown
Toll-5 (tehao)	CG7121	FBgn0026760	2L	13435622	13439333	Signaling	Toll
Toll-6	CG7250	FBgn0036494	3L	15329792	15337417	Signaling	Other/Unknown
Toll-7	CG8595	FBgn0034476	2R	15714083	15720478	Recognition	Cellular; Epithelial; Anti-viral
Toll-9	CG5528	FBgn0036978	3L	20354943	20359824	Signaling	Humoral
Tollo/Toll-8	CG6890	FBgn0029114	3L	15228719	15235932	Signaling	Epithelial
TotB	CG5609	FBgn0038838	3R	16699660	16700244	Effector (Putative)	Other/Unknown
TotC	CG31508	FBgn0044812	3R	16698710	16699302	Effector (Putative)	JAK-STAT
TotF	CG31691	FBgn0044811	2L	19911815	19912377	Effector	Other/Unknown

TABLE S1 (continued).

Name	ID	Fly Base ID	Chromosome	Start	End	Functional Class	Immune Process
TotM	CG14027	FBgn0031701	2L	5329857	5330464	Anti-viral	IMD; JAK-STAT; Anti-viral
TotX	CG31193	FBgn0044810	3R	16730639	16731240	Effector (Putative)	Other/Unknown
TotZ	CG31507	FBgn0044809	3R	16703458	16704085	Effector (Putative)	Other/Unknown
Traf1	CG3048	FBgn0026319	2L	4361925	4380716	Signaling	JNK
Traf6/TRAF2	CG10961	FBgn0265464	X	8048632	8051419	Signaling	Toll; JNK
Transferrin 1	CG6186	FBgn0022355	X	18281117	18284377	Effector	Other/Unknown
Transferrin 3	CG3666	FBgn0034094	2R	12114006	12117374	Effector	Other/Unknown
Transglutaminase	CG7356	FBgn0031975	2L	8011405	8026898	Effector	Coagulation
trpml	CG8743	FBgn0262516	3L	19707518	19711428	Signaling	Phagocytosis
Tsf2	CG10620	FBgn0036299	3L	12523536	12526824	Effector	Other/Unknown
Tsg101	CG9712	FBgn0036666	3L	16835394	16837685	Signaling	JAK-STAT; Phagocytosis
tube	CG10520	FBgn0003882	3R	213464	215535	Signaling	Toll
Turandot A (TotA)	CG31509	FBgn0028396	3R	16696758	16697422	Effector (Putative)	IMD; JAK-STAT
Uba2	CG7528	FBgn0029113	3L	8118427	8120918	Signaling	Toll
UEV1a	CG10640	FBgn0035601	3L	5355984	5359325	Signaling	IMD
Ulp1	CG12359	FBgn0027603	X	19096136	19102220	Signaling	Toll
upd2	CG5988	FBgn0030904	X	18134687	18137317	Signaling	JAK-STAT; Epithelial
upd3	CG33542	FBgn0053542	X	18171266	18178629	Signaling	JAK-STAT; Epithelial
ush	CG2762	FBgn0003963	2L	476220	540560	Signaling	Toll; Cellular
Vago	CG2081	FBgn0030262	X	10983323	10984140	Anti-viral	Anti-viral
vav	CG7893	FBgn0040068	X	19155853	19167313	Signaling	Encapsulation
Victoria (TotE)	CG33117	FBgn0053117	2L	19912938	19913516	Effector	Other/Unknown
vig	CG4170	FBgn0024183	2L	15062931	15070316	Anti-viral	Anti-viral
vir-1	CG31764	FBgn0043841	2L	12403270	12423439	Anti-viral	JAK-STAT; Anti-viral
Vps16A	CG8454	FBgn0261241	3R	5092741	5095946	Other	Phagocytosis
Vps16B	CG18112	FBgn0039702	3R	25640643	25642173	Other	Phagocytosis
Vps28	CG12770	FBgn0021814	2R	3966374	3967546	Other	Phagocytosis
Vps33B	CG5127	FBgn0039335	3R	21299212	21301478	Other	Phagocytosis
Vps4	CG6842	FBgn0027605	X	17801725	17805130	Other	Phagocytosis
wengen	CG6531	FBgn0030941	X	18518075	18528941	Signaling	JNK
wisp	CG15737	FBgn0260780	X	11787976	11793953	Signaling	Toll
wntD	CG8458	FBgn0038134	3R	9117774	9118920	Signaling	Toll
Wrinkled	CG5123	FBgn0003997	3L	18160842	18178741	Signaling	IMD
Yantar	CG18426	FBgn0021895	2R	19734684	19739003	Signaling	Cellular
yellow-f	CG18550	FBgn0041710	3R	8817171	8818923	Effector	PO-Melanization
yellow-f2	CG8063	FBgn0038105	3R	8819117	8820869	Effector	PO-Melanization
yin	CG2913	FBgn0265575	X	3756999	3765972	Other	Phagocytosis
Zfh1	CG1322	FBgn0004606	3R	26591648	26614205	Signaling	Cellular
Zir	CG11376	FBgn0031216	2L	94739	102086	Signaling	Encapsulation

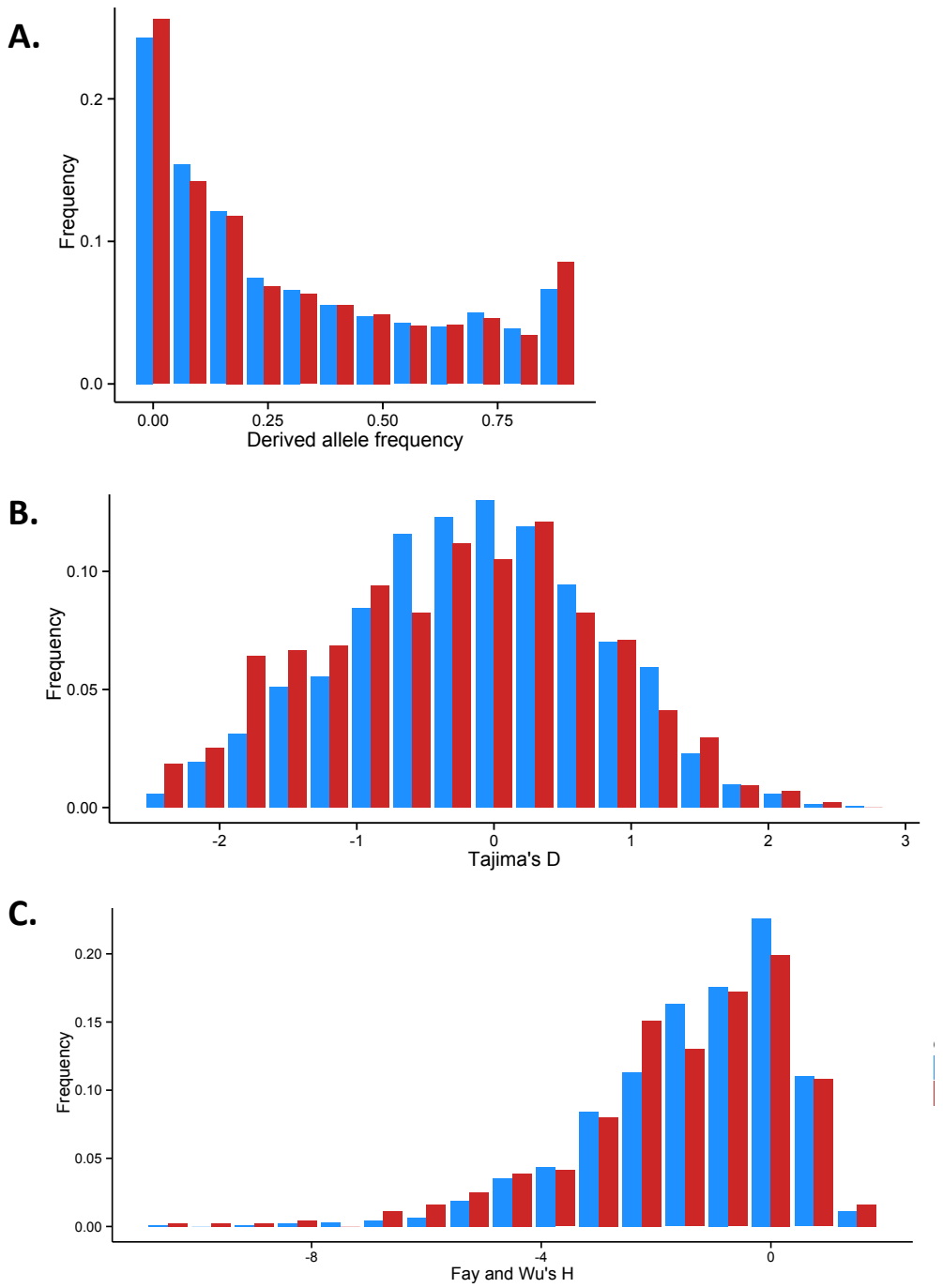


FIGURE S1. Evidence of selection on genes involved in encapsulation: Netherlands.
(A) In the Netherlands, encapsulation genes (red) contain a higher proportion of high frequency derived SNPs relative to control genes (blue; K-S test, $P < 0.00001$, after Bonferroni correction). Encapsulation genes also trend towards more extreme negative values of (B) Tajima's D and (C) Fay and Wu's H although the significance of these patterns do not survive multiple testing correction. (Tajima's D: Mann-Whitney U test, $P = 0.0836$; Fay and Wu's H: Mann-Whitney U test, P = N.S. Significance values are reported after Bonferroni corrected for multiple testing.) Tajima's D and Fay and Wu's H were calculated within 1-kb windows.

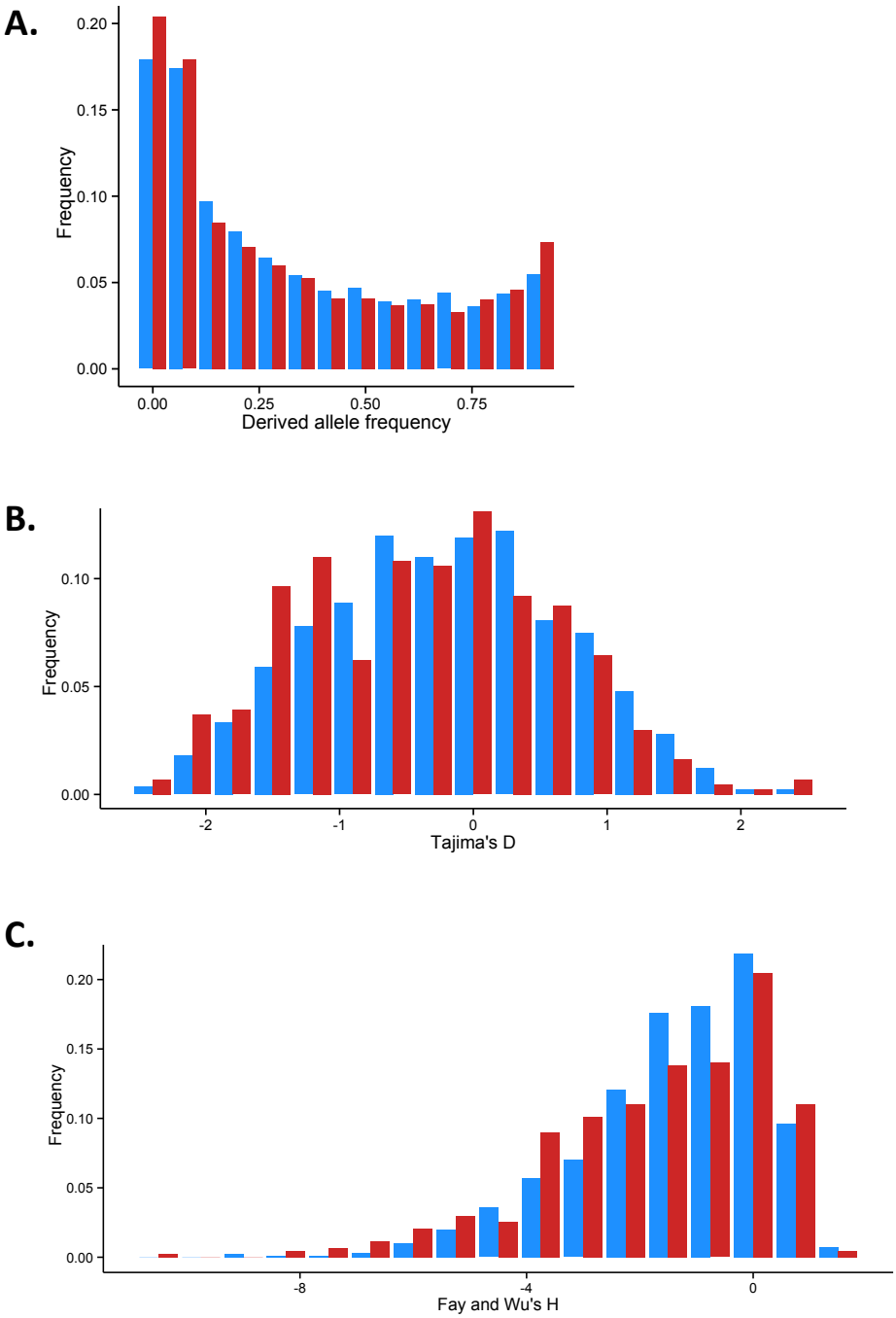


Figure S2. Evidence of selection on genes involved in encapsulation: Beijing.
(A) In Beijing, encapsulation genes (red) contain a higher proportion of high frequency derived SNPs relative to control genes (blue; K-S test, P = 0.00979, after Bonferroni correction). Encapsulation genes also show more extreme negative values of (B) Tajima's D and (C) Fay and Wu's H (Tajima's D: Mann-Whitney U test, P = 0.0125; Fay and Wu's H: Mann-Whitney U test, P = N.S.). Significance values are Bonferroni corrected for multiple testing. Tajima's D and Fay and Wu's H were calculated within 1-kb windows.

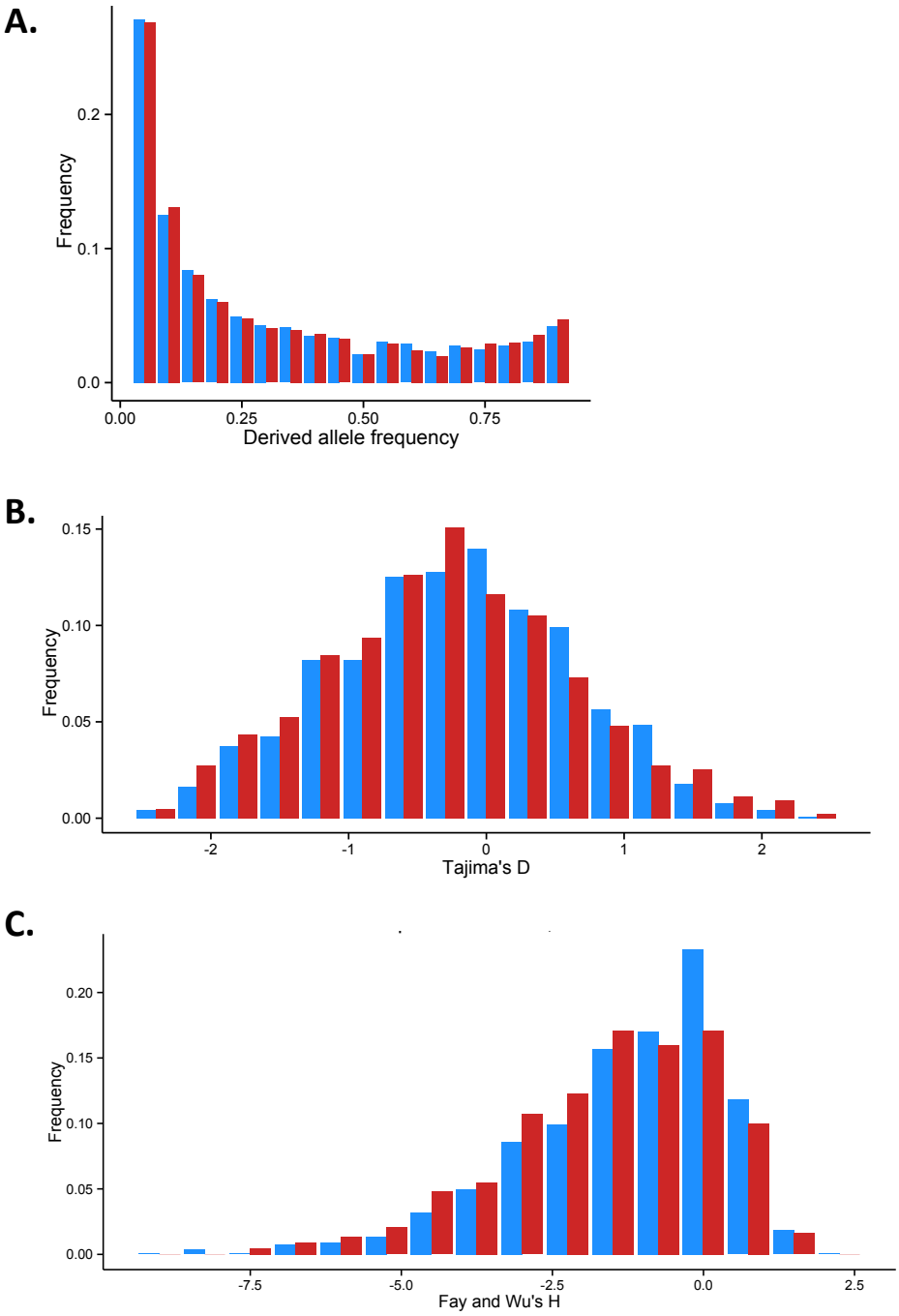


FIGURE S3. Evidence of selection on genes involved in encapsulation: Ithaca, NY.
(A) In Ithaca, NY, encapsulation genes (red) trend towards having a higher proportion of high frequency derived SNPs relative to control genes (blue) although the significance of this pattern does not survive correction for multiple testing (K-S test, P = 0.0777, after Bonferroni correction). Encapsulation genes also show more extreme negative values of (B) Tajima's D and (C) Fay and Wu's H (Tajima's D: Mann-Whitney U test, P = N.S.; Fay and Wu's H: Mann-Whitney U test, P = 0.0479). Significance values are Bonferroni corrected for multiple testing. Tajima's D and Fay and Wu's H were calculated within 1-kb windows.